ADAPTIVE LEARNING: ALGORITHMS AND COMPLEXITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Dylan James Foster

May 2019

ADAPTIVE LEARNING:
ALGORITHMS AND COMPLEXITY

Dylan James Foster, Ph.D.
Cornell University 2019

Recent empirical success in machine learning has led to major breakthroughs in application domains including computer vision, robotics, and natural language processing. There is a chasm between theory and practice here. Many of the most impressive practical advances in learning rely heavily on parameter tuning and domain-specific heuristics, and the development effort required to deploy these methods in new domains places a great burden on practitioners. On the other hand, mathematical theory of learning has excelled at producing broadly applicable algorithmic principles (stochastic gradient methods, boosting, SVMs), but tends to lag behind in state-of-the-art performance, and may miss out on practitioners' intuition. Can we distill our collective knowledge of "what works" into learning procedures that are general-purpose, yet readily adapt to problem structure in new domains?

We propose to bridge the gap and get the best of both worlds through *adaptive learning*: Learning procedures that go beyond the worst case and automatically exploit favorable properties of real-world instances to get improved performance.

The aim of this thesis is to develop adaptive algorithms and investigate their limits, and to do so in the face of real-world considerations such as computation, interactivity, and robustness. In more detail, we:

1. introduce formalism to evaluate and assert optimality of adaptive learning procedures.

2. develop tools to prove fundamental limits on adaptivity.

3. provide efficient and adaptive algorithms to achieve these limits.

In classical statistical decision theory, learning procedures are evaluated by their worst-case performance (e.g., prediction accuracy) across all problem instances. Adaptive learning evaluates performance not just worst case, but in the *best case* and in between. This necessitates the development of new statistical and information-theoretic ideas to provide instance-dependent performance guarantees, as well as new algorithmic and computational principles to derive efficient and adaptive algorithms.

The first major contribution this thesis makes concerns sequential prediction, or online learning. We prove the equivalence of adaptive algorithms, probabilistic objects called martingale inequalities, and geometric objects called Burkholder functions. We leverage the equivalence to provide:

1. a theory of learnability for adaptive online learning.

2. a unified algorithm design principle for adaptive online learning.

The equivalence extends the classical Vapnik-Chervonenkis theory of (worst-case) statistical learning to adaptive online learning. It allows us to derive new learning procedures that efficiently adapt to problem structure, and serves as our starting point for investigating adaptivity in real-world settings.

In many modern applications, we are faced with data that may be streaming, non-i.i.d., or simply too large to fit in memory. In others, we may interact with and influence the data generating process through sequential decisions. Developing adaptive algorithms for these challenges leads to fascinating new questions. Must we sacrifice adaptivity to process and make predictions from data as it arrives in a stream? Can we adapt while balancing exploration and exploitation?

Major contributions this thesis makes toward these questions include:

- We introduce a notion of "sufficient statistics" for online learning and show that this definition leads to adaptive algorithms with low memory requirements.
- We develop large scale optimization algorithms for learning that adapt to problem structure via automatic parameter tuning, and characterize their limits.
- We give adaptive algorithms for interactive learning/sequential decision making in contextual bandits, a simple reinforcement learning setting. Our main result here is a new *margin theory* paralleling that of classical statistical learning.
- We provide robust sequential prediction algorithms that obtain optimal instance dependent performance guarantees for statistical learning, yet make *no assumptions on the data generating process*. We then characterize their limits.
- We design algorithms that adapt to model misspecification in the ubiquitous statistical task of logistic regression. Here we give a new improper learning algorithm that attains a doubly-exponential improvement over sample complexity lower bounds for proper learning. This resolves a COLT open problem of McMahan and Streeter (2012), as well as two open problems related to adaptivity in bandit multiclass classification (Abernethy and Rakhlin, 2009) and online boosting (Beygelzimer et al., 2015).

BIOGRAPHICAL SKETCH

Dylan Foster was born in Phoenix, Arizona in 1992. He obtained his B.S. and M.S. in electrical engineering from the University of Southern California in 2014, and began his Ph.D. in computer science at Cornell University under the supervision of Karthik Sridharan the same year. While at Cornell, he was a recipient of the NDSEG Ph.D. fellowship, Facebook Ph.D. fellowship, and Best Student Paper Award at COLT 2018. After graduating from Cornell, he will join the Massachusetts Institute of Technology as a postdoctoral researcher at the MIT Institute for Foundations of Data Science.

To my family.

ACKNOWLEDGEMENTS

Karthik Sridharan has been everything I could ask for as an advisor, and has been instrumental in molding me into who I am as a researcher, influencing both how I choose problems and how I attack them. His relentless creativity and passion for deep and fundamental problems is truly inspiring, and I am fortunate to have spent the last four years working with him and learning from him. Karthik has been extremely generous with his time throughout my PhD and is always eager to explain new concepts. This was invaluable early on when I was making the switch into theoretical research. Beyond this, Karthik is thoughtful, kind, and easy-going.

I have worked with many outstanding collaborators over the last four years. In roughly chronological order: Karthik, Sasha Rakhlin, Daniel Reichman, Zhiyuan Li, Thodoris Lykouris, Éva Tardos, Satyen Kale, Mehryar Mohri, Peter Bartlett, Matus Telgarsky, Haipeng Luo, Alekh Agarwal, Miro Dudík, Rob Schapire, Ayush Sekhari, and Akshay Krishnamurthy. I am grateful to all of them for their patience, encouragement, and friendship. The contents of this thesis have benefited especially from long term collaborations with Sasha, Satyen, Mehryar, Haipeng and Akshay.

My committee members, Bobby Kleinberg, Éva Tardos, and Kilian Weinberger, were great sources of advice throughout my PhD.

I had three productive internships over the course of my PhD. These contributed in part to key results in this thesis.

First, I thank Sanjiv Kumar for hosting me in his group at Google Research NYC in summer 2016, which led to my collaboration with Satyen and Mehryar. I had a great time talking to the other researchers and visitors in the group, including Elad Hazan, Corinna Cortes, Mario Lucic, Bo Dai, Felix Yu, and Dan Holtmann-Rice.

Second, I thank Rob Schapire and Miro Dudík for hosting me at Microsoft Research NYC in summer 2017. While there, I had the fortunate to collaborate with Akshay Krishnamurthy, Alekh Agarwal, and Haipeng Luo, and I enjoyed the company of the broader MSR machine learning crew, including John Langford, Hal Daume III, and the other ML interns: Christoph Dann, Hoang Le, Alberto Bietti.

Third, I thank Vasilis Syrgkanis for hosting me at Microsoft Research New England in summer 2018. Vasilis is an amazing collaborator with endless energy. I also enjoyed chatting with Nishanth Dikkala, Mert Demirer, Nilesh Tripuraneni, Khashayar Khosravi, Lester Mackey, Greg Lewis, Nika Haghtalab, Ohad Shamir, Adam Kalai, and Jennifer Chayes.

I had the fortune to spend spring of 2017 at the Simons Institute at UC Berkeley for their Foundations of Machine Learning program. Beyond the content of the program itself, I fondly remember all the late night runs to La Burrita and Top Dog with Matus and Andrej.

The students in the Computer Science department at Cornell have been great company. While there are far too many to list, I have particularly enjoyed hanging out with Jonathan Shi, Rediet Abebe, Jack Hessel, Ayush Sekhari, and Eric Lee, and Stephen Mcdowell. I thank all the members of the theory lab, including Jonathan Shi, Sam Hopkins, Rediet Abebe, Thodoris Lykouris, Rad Niazadeh, Pooya Jalaly, Hedyeh Beyhaghi, Manish Raghavan, Yang Yuan, Ayush Sekhari, Michael Roberts, and Ramhtin Rotabi.

I had great housemates throughout my time in Ithaca. Stephen, Sean, Eric, and Ayush: Thank you all for reminding me to goof off once in a while.

Finally, I thank my family: My parents Kim and Susan Foster and my sister Jamie. None of this would have been possible without their constant support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

Overview

CHAPTER 1

INTRODUCTION

Essentially, all models are wrong, but some are useful.

George E.P. Box (1987)

In the last decade, machine learning has been a driving force behind core advances in computer vision (LeCun et al., 2015), robotics (Lillicrap et al., 2015), natural language processing and machine translation (Bahdanau et al., 2014), control and planning (Mnih et al., 2015; Silver et al., 2016), computational biology, recommender systems, information retrieval, and beyond. There are many important statistical and algorithmic lessons to be taken from these advances, yet the most impressive achievements—recognizing cats and dogs in photographs or controlling Atari agents—depend on countless hours of parameter tweaking and substantial domain-specific insights. Can we distill our hard-won understanding of what works in the real world into principled machine learning solutions that can be readily deployed in new domains as needed? Can we do so without sacrificing the statistical accuracy and computational efficiency that has made these advances so significant in the first place?

Machine learning, both as a research discipline and as an applied field, can broadly be understood as trying to solve three problems: modeling (What models work well, and where?), evaluation (What do we actually mean when we say a model works well?), and algorithm design (How can we better train models?). Domingos (2012) describes this succinctly as $\mathsf{learning} = \mathsf{representation} + \mathsf{evaluation} + \mathsf{optimization}$. Historically, important progress on these problems has come both from theoretical research ("theory") and empirical research ("practice"). Theory attempts to make progress through improved mathematical understanding, and has contributed general algorithmic principles that have enjoyed widespread adoption, such as boosting (Freund and Schapire, 1996, 1997; Schapire et al., 1997), convex relaxations for high-dimensional statistics (Donoho, 1995; Candès et al., 2006; Candès and Recht, 2009), and stochastic gradient methods for large-scale learning (Bottou and Bousquet, 2008; Shalev-Shwartz et al., 2011). Practice takes an engineering approach and builds real-world learning systems in pursuit of better performance on concrete tasks, and has led both to general principles (e.g., feedfoward neural networks) and domain-specific insights (e.g., feedforward neural networks with *convolutional* layers for vision) (LeCun et al., 1998).

1.1 Adaptive Learning

The aim of this thesis is to develop systematic tools to promote interplay between theory and practice. We introduce general-purpose learning procedures that can be easily deployed across different domains, yet quickly adapt to problem-specific structure to obtain strong performance, and do so provably. We term this type of guarantee *adaptive learning*. In this thesis we:

- provide a formalism to describe and evaluate adaptive learning procedures.
- establish fundamental limits of adaptive learning.
- develop efficient and adaptive algorithms to match these limits.

What problem structure one should adapt to—equivalently, what type of data is "easy" or "nice"—may vary considerably across application domains. A statistician's idea of niceness could be data sparsity, where examples have only a few relevant features in spite of being very high-dimensional, while a researcher applying machine learning to computer vision might imagine that nice examples are those with spatial regularity. To proceed, it will be helpful to expand on the first bullet and give a formal definition of what it means for a learning procedure to be adaptive.

This thesis formalizes adaptive learning through the language of statistical decision theory (Van der Vaart, 2000; Lehmann and Casella, 2006). Imagine that a user would like to predict or estimate something about the true state of the world. We call this state $\mathsf{Unknown}$. The user (or, "learner") does not have direct access to $\mathsf{Unknown}$, but can gather an observable quantity denoted as $\mathsf{Observable}$, which is generated from $\mathsf{Unknown}$. We write this process as "$\mathsf{Observable} \sim \mathsf{Unknown}$". The user feeds $\mathsf{Observable}$ into a learning procedure Alg that uses it to make a decision, and the quality of this decision is measured via the *risk*

$$\mathop{\mathbb{E}}_{\mathsf{Obs.}\sim\mathsf{Unknown}}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})],$$

where $\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})$ is the price Alg pays for making its decision from $\mathsf{Observable}$ when the true state is $\mathsf{Unknown}$.

For image classification, we might imagine that $\mathsf{Unknown}$ is an unknown mapping from feature vectors (images) to labels ("cat" or "dog"), $\mathsf{Observable}$ is a collection of labeled examples, and Error measures accuracy of a classifier trained on these examples using Alg. While simple, this formulation is quite general and includes many tasks beyond classical supervised classification and regression, including unsupervised learning (e.g., mean estimation or dimensionality reduction), hypothesis testing, and even tasks like stochastic optimization and sequential decision making that do not necessarily fall into the i.i.d. paradigm.

There is substantial research across computer science, statistics, information theory, and optimization that develops so-called *worst-case* guarantees on the risk of statistical decision procedures. These results—one may think of probably approximately correct (PAC) learning (Valiant, 1984), Vapnik-Chervonenkis (VC) theory (Vapnik and Chervonenkis, 1971), or statistical minimax theory (Wald, 1939)—typically provide upper bounds of the form

$$\mathop{\mathbb{E}}_{\mathsf{Obs.}\sim\mathsf{Unknown}}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})] \leq \mathbf{C} \quad \forall \text{ unknowns}, \tag{1.1}$$

where "$\forall$ unknowns" denotes that we would like the guarantee to hold regardless of what the true state of the world is. We call the constant $\mathbf{C}$ a *worst-case* or *uniform* bound on the risk because it upper bounds the learning procedure's performance uniformly across all possible unknown states or "instances".

To develop and analyze learning procedures that adapt to problem structure, it will be helpful to have a more refined notion of statistical performance. We would like to evaluate the performance of decision rules not just on their worst-case performance, but on their best-case performance on instances that are particularly nice and, more broadly, on instances across the whole spectrum of niceness.

Our starting point is to assume that the metric through which niceness is quantified is fixed, and then evaluate statistical decision rules based on the extent to which they adapt in accordance with the metric. That is, we take as given a function $\phi(\mathsf{Observable}, \mathsf{Unknown})$ that specifies jointly the niceness of nature ($\mathsf{Unknown}$) and niceness of the observations ($\mathsf{Observable}$). We call such a function ϕ an *adaptive risk bound*, and a learning procedure Alg will be said to *achieve* ϕ if

$$\mathbb{E}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})] \leq \mathbb{E}[\phi(\mathsf{Observable}, \mathsf{Unknown})] \quad \forall \text{ unknowns}. \tag{1.2}$$

We abbreviate $\mathbb{E}_{\mathsf{Observable}\sim\mathsf{Unknown}}$ to $\mathbb{E}$ above and for the remainder of the chapter.

The utility of this formulation is to abstract away the problem of deciding which instances are nice, which we emphasize is inherently subjective (indeed, the "no-free lunch" theorems (Wolpert, 1996) imply that some instances must be difficult for a given learner). As a rule of thumb we will have the following desiderata:

1. $\phi(\mathsf{Observable}, \mathsf{Unknown})$ should be small whenever $\mathsf{Observable}$ and $\mathsf{Unknown}$ are nice.
2. $\phi(\mathsf{Observable}, \mathsf{Unknown})$ should be not much larger than the best uniform bound $\mathbf{C}$ in the worst case.

Examples of Adaptivity So as not to risk becoming too abstract, let us take a moment to sketch how this framework captures some interesting types of problem structure. For concreteness we focus on supervised learning; either classification or regression. Typically the first step in supervised learning is to pick a model, that is, a set of candidate regression functions or classifiers that map features to targets. We do so with the tacit assumption that the model will be a good fit for nature. In basic data analysis tasks one might use a linear model, and in computer vision or machine translation one might use a deep neural network. Once the model is picked, we gather data and feed it into a learning procedure that uses it to find a regression function or classifier that (hopefully) predicts well on future examples.

In this setting adaptivity captures the interaction between the model and nature in a number of familiar ways.

- *Adaptivity to label or target distribution.* Suppose our goal is to learn a binary classifier to distinguish images of cats and dogs, and suppose we have done a very good job of picking our model: One of the classifiers under consideration perfectly separates the examples in our dataset into cats and dogs! Can we exploit this good fortune to achieve strong predictive performance on future examples? In other words, we would like to achieve the adaptive rate

$$\phi(\mathsf{Observable}, \mathsf{Unknown}) = \text{“small if data is separable.”}$$

  Dating back to the Perceptron (Rosenblatt, 1958), such *margin bounds* have been a core object of study throughout the development of statistical learning theory (Vapnik, 1998; Panchenko, 2002). Algorithms that exploit the margin to predict confidently ("maximizing the margin") such as boosting (Schapire et al., 1997) have enjoyed significant practical success, and more recently the margin has also been recognized to play a role in generalization in deep learning (Zhang et al., 2017; Bartlett et al., 2017).

More generally, we may hope for an adaptive rate that smoothly interpolates between the separable and non-separable regimes, even in the presence of possible model mis-specification, e.g. $\phi(\textsf{Observable}, \textsf{Unknown}) = \textsf{"small if target variance is small"}$. Beyond statistical learning, the importance of exploiting low noise or variance in targets has been studied intensely in closely related areas including sequential decision making (e.g., bandits) (Auer et al., 2002a; Audibert and Bubeck, 2010; Goldenshluger and Zeevi, 2013; Bastani and Bayati, 2015), stochastic optimization (Nemirovski et al., 2009; Lan, 2012), and econometric applications such as learning treatment policies (Chernozhukov et al., 2016; Athey and Wager, 2017).

- *Adaptivity to model class structure:* A basic rule of thumb in learning is that the amount of data one must gather to train a model should scale with the model complexity (Friedman et al., 2001). If our goal is to train a very large model class, we may hope that if data is nice we do not pay for the complexity of the full model but instead pay the complexity of a smaller subclass. In particular, if our model decomposes into a sequence of nested models $\textsf{model}(1) \subset \textsf{model}(2) \subset \ldots$, an adaptive learning guarantee may take the form

$$\phi(\textsf{Observable}, \textsf{Unknown}) = \textsf{"small if model}(i) \textsf{ fits nature well, where } i \textsf{ is not too large."}$$

Such adaptivity is the aim of classical statistical task of *model selection* (Mallows, 1973; Akaike, 1974; Massart, 2007). Model selection is a major feature of high-dimensional statistical procedures such as the Lasso (Donoho, 1995; Candès et al., 2006; Candes and Tao, 2007) that perform data-driven selection of features. The basic challenge of model selection—especially at large scale—is pervasive in modern machine learning, with problems such as choosing the best neural network architecture receiving intense interest (Snoek et al., 2012; Zoph and Le, 2016). The task of learning the best parameters for online or stochastic optimization procedures is closely related (McMahan and Abernethy, 2013).

- *Adaptivity to feature distribution.* A final, ubiquitous type of adaptivity is to achieve improved statistical performance when features themselves have extra structure. There are many natural types of structure that features can present, for example

$$\phi(\mathsf{Observable}, \mathsf{Unknown}) = \begin{array}{l} \text{“small if features are sparse,”} \\ \text{“small if features are low-dimensional,”} \\ \text{“small if features lie on a smooth manifold,”} \\ \vdots \end{array}$$

  Many algorithms for supervised learning and statistical inference exploit that niceness in the feature distribution reduces the “effective complexity” of the model class (Bartlett and Mendelson, 2003; Chandrasekaran et al., 2012; Negahban et al., 2012). This type of adaptivity is also closely related to unsupervised learning, especially dimensionality reduction (Roweis and Saul, 2000; Tenenbaum et al., 2000; Belkin and Niyogi, 2003; Candès et al., 2011). In online and stochastic optimization, adaptive methods that exploit feature sparsity and related structure (Duchi et al., 2011; Kingma and Ba, 2015) have had significant practical impact on large-scale learning.

An important takeaway from these examples is that adaptivity is not simply an issue of tightening analysis. Adaptive learning guarantees typically require explicitly adaptive algorithms and, conversely, algorithms designed with the worst case in mind are often conservative in nature.

1.2 The Adaptive Minimax Principle

The examples in the previous section are themselves but a few points living in a vast space of adaptive learning procedures. This thesis develops theoretical tools to explore this space, and to shed light on the common structure shared by these procedures. There are several interesting and practical questions regarding the *tradeoffs* of adaptive learning that we would like to elucidate. Can we guarantee good performance on a given class of nice instances without sacrificing performance on other instances, or does adapting to niceness come with a price? Are some notions of niceness intrinsically at odds with each other? Can all instances be equally nice? Can certain adaptive learning procedures dominate other adaptive procedures?

As a starting point toward answering these questions, this thesis intoduces *minimax analysis* of adaptive learning. The reader may recall that in classical statistical decision theory, Wald's minimax principle (1939) is a criterion for evaluating and comparing the performance of statistical decision procedures. The principle states that statistical decision rules should be evaluated relative to the *minimax risk*, that is, relative to the performance of the decision rule Alg that minimizes

$$\max_{\mathsf{unknowns}} \mathbb{E}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})], \tag{1.3}$$

or in other words, relative to the value

$$\mathcal{V} := \min_{\mathsf{algorithms}} \max_{\mathsf{unknowns}} \mathbb{E}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})]. \tag{1.4}$$

Adaptive minimax analysis generalizes this idea to adaptive learning. The adaptive analogue of the minimax risk (1.4) is what we call the *minimax achievability* for ϕ, defined via

$$\mathcal{V}(\phi) = \min_{\text{algorithms}} \max_{\text{unknowns}} \mathbb{E}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown}) - \phi(\mathsf{Observable}, \mathsf{Unknown})]. \tag{1.5}$$

Minimax achievability has the following immediate interpretation: We are always guaranteed that there exists an algorithm Alg such that

$$\mathbb{E}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})] \leq \mathbb{E}[\phi(\mathsf{Observable}, \mathsf{Unknown})] + \mathcal{V}(\phi) \quad \forall \text{ unknowns},$$

so that if $\mathcal{V}(\phi) \leq 0$ we can conclude the rate ϕ is indeed achievable. Conversely, for any $c < \mathcal{V}(\phi)$ no algorithm can guarantee

$$\mathbb{E}[\mathsf{Error}(\mathsf{Alg}(\mathsf{Observable}), \mathsf{Unknown})] \leq \mathbb{E}[\phi(\mathsf{Observable}, \mathsf{Unknown})] + c \tag{1.6}$$

for all instances, or in other words the rate $\phi(\mathsf{Observable}, \mathsf{Unknown}) + c$ is *never* achievable.

The *adaptive minimax principle* we employ throughout this thesis is to evaluate adaptive learning procedures according to the smallest c for which they guarantee the inequality (1.6) holds for all instances. While simple, the adaptive minimax principle has a powerful consequence: It allows us to assert *optimality* of adaptive learning procedures. In particular, we say that any procedure Alg that guarantees (1.6) with $c = \mathcal{V}(\phi)$ is *optimal for* ϕ.

We briefly remark that there are many other criteria for evaluating statistical decision procedures, notably Bayesian frameworks (Berger, 2013) and their frequentist relatives (Shawe-Taylor et al., 1998; McAllester, 1999). A key difference is that while both Bayesian frameworks and the adaptive framework (through ϕ) incorporate prior knowledge, the adaptive framework is *agnostic* and does not assume any particular model for the world.

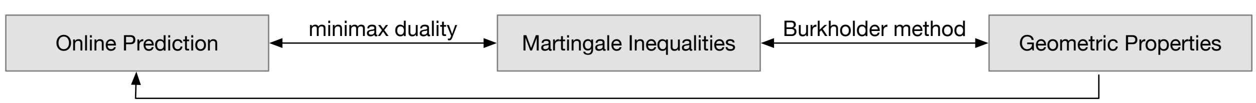
1.2.1 Contribution: Equivalence

This thesis uses the adaptive minimax principle as a starting point to develop an (algorithmic) theory of learnability for adaptive learning in the online prediction (or, online learning) model. The new theory is analogous to the classical PAC or VC theory for statistical learning (Valiant, 1984; Vapnik and Chervonenkis, 1971), but characterizes achievability and rates for *adaptive learning*, and does so in the online setting. It is based on the following *equivalence*:

Adaptive risk bounds are equivalent to mathematical objects called martingale inequalities, which are in turn equivalent to geometric objects called Burkholder functions.

Let us give a bird's-eye view of the result:

Figure 1.1:



Beginning with the notion of minimax achievability, we first show that for any adaptive rate ϕ, achievability is equivalent to a corresponding probabilistic martingale inequality. This is achieved with the help of the minimax theorem. We then turn to the Burkholder method—a tool developed in a series of celebrated works by Donald Burkholder to *certify* martingale inequalities (Burkholder, 1981, 1984, 1986, 1991)—and show equivalence of these martingale inequalities and existence of a special Burkholder (or, "Bellman") function, a purely geometric object. Finally, we use this function for adaptive online prediction, thus completing the circle. The main consequences of this development are:

1. Martingale inequalities characterize the *fundamental limits* of adaptive learning in

an algorithm-independent manner. Consequently, the probabilist's toolbox of tail bounds, maximal inequalities, and so forth may be used to certify the existence or non-existence of algorithms without concern for algorithm design.

2. Once an adaptive learning guarantee is known to be achievable, the geometric certificates (Burkholder functions) provided by the Burkholder method can be exploited to design *efficient* algorithms.

1.3 Adaptive Learning for Real-World Challenges

The improved understanding of adaptivity and adaptive algorithms provided by the equivalence has strong consequences for real-world machine learning and statistics applications. In such applications, learning algorithms must be evaluated under practical considerations; classical (e.g., PAC) statistical risk is not a holistic measure of performance. For example, learning procedures with strong statistical performance may be useless in practice if they are difficult to compute or do not fit in memory. Learning procedures may not be deployed in purely observational settings, but instead may be used to make *decisions* that influence future observations and outcomes (e.g., robotic control).

To leverage the equivalence to design algorithms that are adaptive and *practical*, we make the observation that many of practical constraints can be encoded in the adaptive minimax framework (1.5)—a consequence of its high generality. We outline several examples below.

- *Computation and memory.* Real-world machine learning is concerned not just with statistical performance, but statistical performance subject to the constraint that learning procedures run in a reasonable amount of time, and do so while using a reasonable amount of memory. Issues of computation in learning date back to

Valiant's work on PAC learning (Valiant, 1984), which is concerned with *polynomial time* learnability. In recent decades, a more refined understanding of the interplay between computation and learning has developed, including a limited understanding of fundamental tradeoffs between computation time and statistical efficiency (Decatur et al., 2000; Servedio, 2000; Bottou and Bousquet, 2008; Shalev-Shwartz et al., 2011, 2012; Chandrasekaran and Jordan, 2013; Berthet et al., 2013; Zhang et al., 2014).

One line of research in this thesis develops computationally efficient adaptive algorithms through the *online learning* model. Learning procedures for online learning framework such as stochastic gradient methods offer—both in theory and practice—an effective knob with which to control tradeoffs between computation time and statistical accuracy (Bottou and Bousquet, 2008; Shalev-Shwartz et al., 2011). Moreover, many algorithms developed for online learning enjoy low memory requirements, even though this is not formally part of the model. This thesis includes some new contributions toward making this connection more formal.

This thesis also explores interplay between adaptivity and computation through optimization complexity. Modern machine learning is ripe with (stochastic) optimization problems in which we search for the model that minimizes the (empirical or population) risk or related objectives. A basic unit of computation for such problems is the *oracle complexity* or *information-based complexity* (Nemirovski et al., 1983): Given an oracle that accepts a point and returns the function value, gradient, or other features, how many oracle queries are required to approximately minimize the function to a desired precision. We develop adaptive algorithms in the *online convex optimization* model, which immediately implies (adaptive) upper bounds on the oracle complexity of stochastic optimization problems arising in learning.

- *Interactivity.* Systems in which agents learn to make decisions by sequentially interacting with an unknown environment are becoming increasingly ubiquitous. These range in complexity from content recommendation systems (Li et al., 2010; Agarwal et al., 2016) (agents present decisions such as news articles to users, learn from these decisions, and improve decisions for future users) to reinforcement learning agents for sophisticated human-level control tasks (Mnih et al., 2015; Silver et al., 2016). Closely related are tasks in causal inference and policy learning (Swaminathan and Joachims, 2015; Chernozhukov et al., 2016, 2018; Athey and Wager, 2017). Interactivity, across the complexity spectrum, induces a tradeoff between exploration and exploitation that must be balanced to ensure sample efficient learning.

In this thesis we address adaptivity in interactive learning through the *contextual bandit* model. Contextual bandits generalize the online supervised learning setting to accommodate uncertainty (specifically, partial or incomplete feedback), and have seen successful application in news article recommendation and mobile health (Li et al., 2010; Agarwal et al., 2016; Tewari and Murphy, 2017; Greenewald et al., 2017). From a technical perspective, the contextual bandit model is a good testbed for developing new algorithmic tools for adaptive learning because it is the simplest reinforcement learning setting that embeds the full complexity of statistical learning.

- *Robustness.* Box (1987) writes: "Essentially, all models are wrong, but some are useful." While modeling is intrinsic to learning, it is important to develop learning procedures that give strong guarantees and degrade gracefully when modeling assumptions fail. Indeed, it is widely recognized that improving robustness is essential step toward building learning systems that can be safely deployed in the real world (Kurakin et al., 2017; Biggio and Roli, 2018). In statistical learning, these issues have been explored through the *agnostic PAC* model (Haussler, 1992; Kearns et al., 1994) and the so-called *general setting of learning* (Vapnik, 1995). Robustness has been addressed in parallel throughout the history of statistics, leading to a complementary set of models and principles for robust inference (Box, 1953; Tukey, 1975; Huber, 1981; Hampel et al., 1986).

This thesis explores the interaction between robustness and adaptivity both in agnostic statistical learning, and in the other learning models mentioned thus far (online learning, online convex optimization, contextual bandits), all of which are agnostic in nature and do not assume model correctness.

1.3.1 Contribution: New Adaptive Learning Guarantees

The practical considerations above lead to a number of fascinating challenges when combined with questions of adaptivity. Computationally, can we develop efficient algorithms that adapt to data whenever this is statistically possible? Is it more difficult to adapt when have to make predictions on the fly for data arriving in a stream? How does adaptivity interact with tradeoffs between exploration and exploitation? This is where the tools provided by the equivalence come to help.

We work in four settings—online learning, online optimization, agnostic statistical learning, and contextual bandits—and for each setting identify an important family of adaptive guarantees for which existing theory and algorithms are unsatisfactory. For each such family we comprehensively characterize the fundamental limits on the degree to which this new type of adaptivity can be achieved, and then design efficient algorithms to achieve this limit. The central contributions are:

- We give tools that permit the systematic development of low-memory adaptive algorithms. We show that whenever a given adaptive rate can be expressed in terms of certain "sufficient statistics" of the data sequence, there exists an online learning algorithm that is only required to keep these sufficient statistics in memory.

- We introduce optimal and efficient algorithms that adapt to problem structure in online convex optimization via online parameter tuning, and characterize limits for this type of adaptivity via connections to the theory of model selection in statistical learning.
- We develop robust statistical learning algorithms that adapt to the degree of model misspecification. Specifically, for logistic regression we design a new improper learning algorithm (via online learning techniques) that attains a doubly-exponential improvement over sample complexity lower bounds for proper learning in the misspecified setting, thereby resolving a COLT open problem of McMahan and Streeter (2012). We then use this algorithm to resolve open problems regarding adaptive algorithms for bandit multiclass classification (Abernethy and Rakhlin, 2009) and online boosting (Beygelzimer et al., 2015), and characterize the extent to which this improvement extends to general hypothesis classes.
- We give a general theory for adapting to problem structure via margin ("margin theory") in the contextual bandit setting, and develop efficient algorithms to match the guarantees from this framework. Our margin theory for contextual bandits applies at the same level of generality as the classical margin theory in statistical learning, but applies to much more challenging sequential decision making tasks.
- We introduce new *sequence optimal* algorithms for online supervised learning that adapt to the structure of the feature distribution. These algorithms are always guaranteed to match the best possible performance in the i.i.d. statistical learning setting, yet do so without making any assumptions on the data generating process. We characterize the limits of this type of adaptivity through a new connection between online learning and probability in Banach spaces.

Themes In the classical statistical learning model, work beginning with Vapnik and Chervonenkis (1971), has led to a diverse and extensive collection of adaptive performance guarantees. These guarantees are obtained by simple algorithms, and by and large may be understood as consequences of basic phenomena in empirical process theory (Pollard, 1990). A central theme the reader should keep in mind throughout the thesis is:

When can adaptive guarantees from the classical (i.i.d.) statistical learning setting also be achieved for more challenging (e.g., sequential or interactive) learning settings?

Beyond exploring achievability of different notions of adaptivity in the information-theoretic sense, it is also of central importance to understand how the algorithmic principles change when we move beyond the classical setting. Both issues are addressed throughout this thesis.

A second theme is that all of the algorithms we develop make few or no assumptions on the process by which data is generated. Even though we might imagine that the real world is ripe with niceness and problem structure, we lose little by working in such agnostic learning models precisely because the *adaptive* algorithms we develop can exploit problem structure whenever instances do happen to be nice.

1.4 Organization

In the remainder of this chapter (Section 1.5), we give a preview of the general approach to analyzing adaptive learning through the equivalence framework. Then, in Chapter 2 we develop the minimax analysis of adaptive learning formally, showing how to formulate statistical learning, online learning, contextual bandits, and online and stochastic optimization in the adaptive minimax framework. From here on, the main content of the thesis is broken into two parts.

Part II: Equivalence of Prediction, Martingales, and Geometry In Part II, we introduce the technical tools that form the core of the thesis. We work in the online learning model, and the main development is the equivalence of adaptive learning, martingale inequalities, and Burkholder functions illustrated in Figure 1.1.

In Chapter 4 we present the equivalence in its simplest form, focusing on the online supervised learning setting with linear losses. As a running example, we illustrate the method by developing a new adaptive algorithm for online matrix prediction. In Chapter 5 we present the equivalence in its general form. We also show—via the Burkholder method—how a certain notion of sufficient statistics for online learning leads to low-memory adaptive algorithms. In Chapter 6 we develop generic tools for proving martingale inequalities that arise from the equivalence. We show how adaptive rates in the supervised learning model induce certain "offset" random processes, and that obtaining small upper bounds on these processes is sufficient to demonstrate achievability. We use this approach to recover a number of existing adaptive guarantees, as well as to derive new guarantees.

Part III: New Guarantees for Adaptive Learning In Part III, with the toolbox from Part II in hand, we proceed to develop new types of adaptive learning guarantees for four settings: statistical learning, online learning, contextual bandits, and online and stochastic optimization. For each setting we identify a new notion of adaptivity, characterize the fundamental limits on the degree to which this adaptivity is achieved, and design efficient algorithms to achieve this limit. Chapter 8 develops sequence-optimal online learning algorithms that adapt to the feature distribution, Chapter 9 introduces algorithms for model selection and parameter tuning in online convex optimization, Chapter 10 gives new algorithms that adapt to model misspecification in logistic regression and related problems, and Chapter 11 develops margin theory for contextual bandits.

1.5 Highlight: Achievability and Algorithm Design

We close the introduction by offering a taste of the tools developed in Part II. We focus on a setting that is extremely simple, yet completely free of assumptions—online bit prediction—and show how a result of Cover (1967) completely answers two key questions:

1. What properties of an adaptive rate function ϕ suffice to guarantee that the rate is achievable?
2. When such a rate ϕ is achievable, what algorithm achieves it?

The bit prediction setting is a special case of the *online learning* setting that features prominently in this thesis. The learning process proceeds in n rounds: At each step t, the learner randomly selects a prediction distribution q_t, receives an outcome $y_t \in \{\pm 1\}$, then samples its prediction $\widehat{y}_t \sim q_t$ and suffers the indicator loss $\mathbb{1}\{\widehat{y}_t \neq y_t\}$. In this setting, adaptive rates $\phi(y_{1:n})$ map the bit sequence $y_{1:n} = y_1, \ldots, y_n$ to a risk bound. A rate ϕ is achieved by the learner if

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}\{\widehat{y}_t \neq y_t\}\right] \leq \phi(y_{1:n}) \qquad \text{for every sequence } y_{1:n},$$

where the expectation is taken with respect to the learner's randomness. To formulate the minimax value for this setting, we think of a sequential game between the learner and nature. We imagine that in the worst case, nature is an adversary whose goal is to make the learner's regret to ϕ as large as possible, so that the goal of a minimax optimal learner is to minimize regret against this adversary. At each round the contribution to regret is a $\min\text{-}\max$ problem conditioned on the history so far: The learner chooses q_t to minimize regret given the history, then nature picks a maximally bad value for y_t given the learner's decision, and finally the prediction $\widehat{y}_t$ is sampled from q_t. The process is repeated for all n rounds, giving rise to the following expression for the minimax value:

$$\mathcal{V}(\phi) = \min_{q_1} \max_{y_1} \mathop{\mathbb{E}}_{\widehat{y}_1 \sim q_1} \cdots \min_{q_n} \max_{y_n} \mathop{\mathbb{E}}_{\widehat{y}_1 \sim q_1} \left[\frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} - \phi(y_{1:n}) \right].$$

So, for what functions ϕ does there exist a strategy for the learner such that this inequality holds (i.e. $\mathcal{V}(\phi) \leq 0$)? Since the adaptive risk inequality is required to hold for *every* sequence $y_{1:n}$, we are free to try some examples to deduce the important properties of ϕ. As a particular choice, let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of independent *Rademacher random variables*, i.e. fair coin flips in $\{\pm 1\}$, and choose $y_t = \epsilon_t$. Since the learner's strategy at time t only depends on $\epsilon_1, \ldots, \epsilon_{t-1}$, it is easy to see that $\mathbb{E}[\mathbb{1}\{\widehat{y}_t \neq y_t\}] = \frac{1}{2}$ for any learner. Since this holds at each round, we conclude that a *necessary* condition for achievability is that

$$\mathop{\mathbb{E}}_{\epsilon}[\phi(\epsilon_{1:n})] \geq \frac{1}{2}. \tag{1.7}$$

This condition is necessary, but is it sufficient? Suppose that ϕ is additionally *stable*, in the sense that

$$|\phi(\epsilon_1, \ldots, \epsilon_t, \ldots, \epsilon_n) - \phi(\epsilon_1, \ldots, \epsilon_t', \ldots, \epsilon_n)| \leq \frac{1}{n}$$

for all sequences $\epsilon_{1:n}$, all choices for ϵ_t', and all times t. Cover's result is that in this case, the answer is *yes*.

Lemma 1 (Cover (1967))**.** Let ϕ be any stable adaptive rate function. Then ϕ is achievable *if and only if* $\mathbb{E}_{\epsilon}[\phi(\epsilon_{1:n})] \geq \frac{1}{2}$. Furthermore, any rate ϕ satisfying this condition is achieved by the algorithm that chooses q_t to be the unique distribution over $\{\pm 1\}$ with mean

$$\mu_t = n \cdot \mathop{\mathbb{E}}_{\epsilon_{t+1:n}} [\phi(y_1, \ldots, y_{t-1}, +1, \epsilon_{t+1}, \ldots, \epsilon_n) - \phi(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n)]. \tag{1.8}$$

Cover's result is proved through a potential function argument, which is a recurring theme in this thesis. The characterization has two favorable properties:

1. The condition for achievability is *algorithm-independent*. Checking for existence of a prediction strategy that achieves ϕ is as simple as checking the probabilistic inequality $\mathbb{E}_{\epsilon}[\phi(\epsilon_{1:n})] \geq \frac{1}{2}$.
2. It admits an explicit algorithm. That is, once a rate ϕ is known to be achievable, we can efficiently compute the strategy that obtains ϕ and use it to make predictions.[1]

Cover's characterization is quite elegant and will serve as an inspiration for our results going forward, but it has a number of insufficiencies that must be addressed if we wish to apply the framework to solve real-world learning challenges. To note a few:

- The learning problem to which the characterization applies does not have covariates or contexts. This prevents it from being applied to basic classification and regression tasks.
- The learner's decision space $\{\pm 1\}$ is quite simple; to develop adaptive algorithms for, e.g. optimization, we should accomodate rich output spaces, such as subsets of $\mathbb{R}^d$ or even infinite-dimensional Banach spaces.
- The characterization only applies to the classification loss. To accommodate standard problems in learning and statistics, we would like to handle other losses, such as the square loss, logistic loss, and so forth. Characterizing the correct statistical complexity and developing optimal algorithms for general losses is far from trivial, even in the case of uniform (non-adaptive) rates.

[1] It is straightforward to show via concentration that the expectation in the strategy's definition can be approximated arbitrarily well with polynomially many samples. This strategy achieves ϕ up to an arbitrarily small additive constant.

- In the supervised learning problems, adaptive rates typically incorporate regret against a benchmark class of models $\mathcal{F}$. For example, we might have

$$\phi(x_{1:n}, y_{1:n}) = \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \mathcal{B}(x_{1:n}, y_{1:n}),$$

  where $\mathcal{B}$ is another function that we refer to as an adaptive bound on the *regret* to $\mathcal{F}$. What properties of $\mathcal{F}$ influence achievability? Are the requirements on ϕ more stringent when $\mathcal{F}$ is a class of neural networks than when it is a class of linear functions? In the case of uniform rates ($\mathcal{B}$ is constant), this question is addressed in a line of work beginning with Rakhlin et al. (2010); we extend this to handle adaptivity.

- The result is specialized to "full information", wherein the learner completely observes the feedback chosen by nature. Can we obtain similar characterizations for the complexity of adaptive learning when the feedback is only partially observed? This is the essential difficulty of contextual bandits.

These questions are far from trivial, and the main results in this thesis may be understood as answering them with varying levels of completeness.

Proof of Lemma 1. We must prove that $\mathbb{E}_{\epsilon}[\phi(\epsilon_{1:n})] \geq \frac{1}{2}$ and stability are sufficient for achievability, and that the strategy q_t achieves ϕ under these conditions.

Let $\mathbf{U}_t(y_1, \ldots, y_t) = \frac{n-t}{2n} - \mathbb{E}_{\epsilon_{t+1:n}} \phi(y_1, \ldots, y_t, \epsilon_{t+1}, \ldots, \epsilon_n)$. Then clearly it holds that

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} - \phi(y_1, \ldots, y_n) \leq \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} + \mathbf{U}_n(y_1, \ldots, y_n).$$

To prove the result, it suffices to show inductively that for each $1 \leq t \leq n$, playing the prescribed strategy ensures

$$\mathbb{E}_{\widehat{y}_t \sim q_t}\left[\frac{1}{n}\mathbb{1}\{\widehat{y}_t \neq y_t\} + \mathbf{U}_t(y_1, \ldots, y_t)\right] \leq \mathbf{U}_{t-1}(y_1, \ldots, y_{t-1}),$$

for any outcome y_t. This implies that the strategy guarantees

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}\{\widehat{y}_t \neq y_t\} - \boldsymbol{\phi}(y_1,\ldots,y_n)\right] \leq \mathbf{U}_0(\cdot),$$

and we have $\mathbf{U}_0(\cdot) = \frac{1}{2} - \mathbb{E}_\epsilon\, \boldsymbol{\phi}(\epsilon_1,\ldots,\epsilon_n) \leq 0$ under the assumption that $\mathbb{E}_\epsilon[\boldsymbol{\phi}(\epsilon_{1:n})] \geq \frac{1}{2}$.

We proceed with the inductive proof. Since q_t is a distribution over $\{\pm 1\}$, it can be parameterized by its mean $\mu_t \in [-1,+1]$. With this parameterization, we have $\mathbb{E}_{\widehat{y}_t \sim q_t}\left[\frac{1}{n}\mathbb{1}\{\widehat{y}_t \neq y_t\}\right] = \frac{(1-\mu_t y_t)}{2n}$. Consequently, the minimax value at time t can be written

$$\min_{\mu_t}\max_{y_t\in\{\pm 1\}}\left[\frac{(1-\mu_t y_t)}{2n} + \mathbf{U}_t(y_1,\ldots,y_t)\right]$$

We choose μ_t so that the value inside the brackets is constant regardless of the outcome y_t, i.e. to guarantee

$$\frac{(1-\mu_t)}{2n} + \mathbf{U}_t(y_1,\ldots,y_{t-1},+1) = \frac{(1+\mu_t)}{2n} + \mathbf{U}_t(y_1,\ldots,y_{t-1},-1),$$

which results in the strategy $\mu_t = n\cdot(\mathbf{U}_t(y_1,\ldots,y_{t-1},+1) - \mathbf{U}_t(y_1,\ldots,y_{t-1},-1))$. The stability property implies that this choice satisfies $\mu_t \in [-1,+1]$, and by direct calculation it is seen that we indeed have

$$\max_{y_t\in\{\pm 1\}}\left[\frac{(1-\mu_t y_t)}{2n} + \mathbf{U}_t(y_1,\ldots,y_t)\right] = \mathbf{U}_{t-1}(y_1,\ldots,y_{t-1}).$$

□

For further results regarding Cover's characterization we refer the reader to Rakhlin and Sridharan (2016b).

1.6 Bibliographic Notes

The results in Chapter 2 and Part II are based on joint work with Alexander Rakhlin and Karthik Sridharan in Foster et al. (2015), Foster et al. (2017b), and Foster et al. (2018c).

From Part III, Chapter 8 is also based on Foster et al. (2017b). Chapter 9 is based on a joint work with Satyen Kale, Mehryar Mohri, and Karthik Sridharan (Foster et al., 2017a). Chapter 10 is based on a joint work with Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan (Foster et al., 2017a). Chapter 11 is based on a joint work Akshay Krishnamurthy (Foster and Krishnamurthy, 2018).

1.7 Notation

General Notation $\mathbb{1}\{\mathcal{E}\}$ will denote the indicator for a meaurable event $\mathcal{E}$, and $\mathbb{P}\{\mathcal{E}\}$ will denote the probability of the event when the measure is clear from context. $\mathbb{E}$ will denote expectation. When P is a probability distribution and X is a formal variable, the notation "$X \sim P$" will be interpreted to mean "X is distributed according to P."

The notation $\sigma(X)$ will denote the Borel σ-algebra for a random variable X.

We define

$$\operatorname{sgn}(x) = \begin{cases} 1, & x > 0. \\ 0, & x = 0. \\ -1, & x < 0. \end{cases}$$

For an integer $k \in \mathbb{N}$ we define $[k] = \{1, \ldots, k\}$. For scalars $a, b \in \mathbb{R}$ we adopt the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

We use $a := b$ to mean "a is defined to be equal to b" and likewise use $a =: b$ to mean "b is defined to be equal to a."

Δ_d will denote the simplex in d dimensions. More generally, we use Δ_A or $\Delta(A)$ to denote the set of all Borel probability measures on the set A.

Asymptotic Notation For functions $f, g : \mathbb{R}^d \to \mathbb{R}$, we say $f \in O(g)$ if there exists a constant C such for all $\mathbb{R}^d$-valued sequences $(\boldsymbol{\alpha}^n)_{n\geq 1}$ with $\lim_{n\to\infty} \boldsymbol{\alpha}_i^n \to \infty$ for all i,

$$\limsup_{n\to\infty} \frac{f(\boldsymbol{\alpha}^n)}{g(\boldsymbol{\alpha}^n)} \leq C.$$

Likewise, we say $f \in \Omega(g)$ if for all such sequences,

$$\liminf_{n\to\infty} \frac{f(\boldsymbol{\alpha}^n)}{g(\boldsymbol{\alpha}^n)} \geq C.$$

We say $f \in \tilde{O}(g)$ and $f \in \tilde{\Omega}(g)$ if $f \in O(g \cdot \operatorname{polylog}(g))$ and $f \in \Omega(g/\operatorname{polylog}(g))$ respectively.

Analysis Throughout this thesis $\|\cdot\|$ will denote a norm and $\|\cdot\|_\star$ will denote the dual. Specific norms include the ℓ_p norms, denoted $\|\cdot\|_p$, the Schatten p-norms, denoted $\|\cdot\|_{S_p}$, the spectral norm $\|\cdot\|_\sigma$, and the nuclear norm $\|\cdot\|_\Sigma$. B_p^d will denote the d-dimension unit ℓ_p ball.

We will use the notation $(\mathfrak{B}, \|\cdot\|)$ to denote a Banach space $\mathfrak{B}$ equipped with norm $\|\cdot\|$, and will let $(\mathfrak{B}^\star, \|\cdot\|_\star)$ denote the dual space. When $x \in \mathfrak{B}$ and $y \in \mathfrak{B}^\star$, $\langle y, x\rangle$ denotes the dual pairing, which coincides with the inner product when $\mathfrak{B}$ is a Hilbert space.

Let $\mathbb{S}^d$ denote the set of symmetric matrices in $\mathbb{R}^{d\times d}$, $\mathbb{S}_+^d$ denote the set of positive-semidefinite (psd) matrices, and $\mathbb{S}_{++}^d$ denote the set of positive-definite matrices. For compatible matrices A and B, $\langle A, B\rangle = \mathsf{tr}(AB^\top)$ is the standard matrix inner product.

A twice-differentiable function $f : \mathfrak{B} \to \mathbb{R}$ is said to be β-smooth with respect to $\|\cdot\|$ if its gradient satisfies $\|\nabla f(x) - \nabla f(y)\|_\star \leq \beta\|x - y\|$. We will use the phrase "smooth norm" to refer to any norm for which the function $\Psi(x) = \frac{1}{2}\|x\|^2$ is β-smooth with respect to $\|\cdot\|$. This is equivalent to the statement that the following inequality holds for all $x, y \in \mathfrak{B}$: $\Psi(y) \leq \Psi(x) + \langle \nabla\Psi(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2$.

A Banach space $(\mathfrak{B}, \|\cdot\|)$ is said to be $(2, D)$-smooth if for all $x, y \in \mathfrak{B}$ (Pinelis, 1994),

$$\|x+y\|^2 + \|x-y\|^2 \leq 2\|x\|^2 + 2D^2\|y\|^2.$$

From this definition it is seen that any Banach spaces with a β-smooth norm has the $(2, \sqrt{\beta/2})$-smoothness property.

A space $(\mathfrak{B}, \|\cdot\|)$ is said to have martingale type 2 with constant β if there exists some $\Psi : \mathfrak{B} \to \mathbb{R}$ such that $\frac{1}{2}\|x\|^2 \leq \Psi(x)$, Ψ is β-smooth with respect to $\|\cdot\|$, and $\Psi(0) = 0$ (Pisier, 1975).

For a function $f : \mathcal{X} \to \mathbb{R}$, we let $f^\star$ denote the Fenchel dual, i.e. $f^\star(y) = \sup_{x \in \mathcal{X}}[\langle y, x \rangle - f(x)]$.

Martingales Let $(X_t)_{t \geq 1}$ be a sequence of real- or $\mathfrak{B}$-valued random variables adapted to a filtration $(\mathcal{F}_t)_{t \geq 0}$. The sequence is said to be *martingale* if

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = X_{t-1} \quad \forall t,$$

and is said to be a *martingale difference sequence* (MDS) if

$$\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0 \quad \forall t.$$

Let $(\epsilon_t)_{t \geq 1}$ be a sequence of Rademacher random variables. A *dyadic* martingale difference sequence is a MDS adapted to the filtration $\mathcal{F}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)$. Any dyadic MDS can be written as

$$X_t = \epsilon_t \cdot \boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}),$$

where $\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1})$ is a *predictable process*.

Miscellaneous Learning Notation We will frequently use the notation $x_{1:n} = x_1, \ldots, x_n$ to refer to a list of examples. When the elements of such a list are vectors in $\mathbb{R}^d$, we will use

$x_t[i]$ to denote the tth vector's ith coordinate. We make reference to the following standard loss functions.

- Indicator loss/zero-one loss: $\ell(\widehat{y}, y) = \mathbb{1}\{\widehat{y} \neq y\}$.
- Absolute loss: $\ell(\widehat{y}, y) = |\widehat{y} - y|$.
- Square loss: $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$.
- Logistic loss: $\ell(\widehat{y}, y) = \log\left(1 + e^{-\widehat{y}y}\right)$.

CHAPTER 2

LEARNING MODELS AND ADAPTIVE MINIMAX FRAMEWORK

A central aim of this thesis is to give a unified formalism for analyzing adaptive learning guarantees in real-world settings. This section lays the groundwork for this approach by developing the adaptive minimax analysis framework outlined in the introduction formally. We instantiate the general framework for learning settings that feature throughout the thesis: statistical learning, online learning, online optimization, and contextual bandits.

2.1 Adaptive Minimax Value

We work in the language of statistical decision theory (Van der Vaart, 2000; Lehmann and Casella, 2006). The class of possible instances in nature is described by a set distributions $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ over a domain $\mathcal{S}$, parameterized by some set Θ (e.g., for mean estimation, $\mathcal{P}$ could be a set of gaussian distributions with Θ describing the set of means). Nature selects an element $\theta \in \Theta$, and the learner receives a sample $S \sim P_\theta$. Letting $\widehat{\Theta}$ denote the set of *decisions*, the learner outputs a (potentially randomized) decision function $\widehat{\theta} : \mathcal{S} \to \widehat{\Theta}$. For a fixed *loss* or *risk* functional $L : \widehat{\Theta} \times \Theta \to \mathbb{R}$, the learner's expected risk is measured via

$$\mathop{\mathbb{E}}_{P_\theta}\Big[L(\widehat{\theta}(S), \theta)\Big]. \tag{2.1}$$

As in the introduction, we let an *adaptive rate functional* $\phi : \mathcal{S} \times \Theta \to \mathbb{R}$ be given, and evaluate the learner's risk relative to ϕ, i.e.

$$\mathop{\mathbb{E}}_{P_\theta}\Big[L(\widehat{\theta}(S), \theta) - \phi(S, \theta)\Big].$$

If this quantity is bounded by zero, the functional ϕ (or "rate", for short) is said to be *achieved* by the rule $\widehat{\theta}$. Note that the rate ϕ captures both niceness in the instance P_θ and niceness of the outcome S itself.

The risk relative to the rate ϕ immediately lends itself to minimax analysis, thereby extending Wald's minimax principle (Wald, 1939) to incorporate adaptivity. This is formalized by the *minimax achievability* for ϕ, defined via

$$\mathcal{V}(\phi) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{P_\theta}\Big[L(\widehat{\theta}(S), \theta) - \phi(S, \theta)\Big]. \tag{2.2}$$

All learning models studied in this thesis share the following structure, which is a special case of the general decision setup (2.1): The observation S will be a sequence of examples of the form $z_1, \ldots, z_n$ with each z_t belonging to some set $\mathcal{Z}$. Individual examples may be drawn i.i.d. (statistical learning) or selected interactively based on the learner's decisions (online learning). The learner makes decisions $\widehat{y}$ belonging to a set $\mathcal{D}$, and the loss for a given prediction-example pair will be $\boldsymbol{\ell}(\widehat{y}, z)$, where $\boldsymbol{\ell} : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$. The final metric through which performance is measured is the value of the function $\boldsymbol{\ell}$ (either empirical or expected, depending on the setting) relative to an adaptive rate ϕ.

Warmup: PAC Learning and Statistical Estimation We first consider a setting that encompasses classical PAC learning (Valiant, 1984), as well as the basic statistical task of statistical estimation with a well-specified model (e.g. Tsybakov (2008)). We take S to be a collection of examples $\{(x_t, y_t)\}_{t=1}^n$ in $\mathcal{X} \times \mathcal{Y} =: \mathcal{Z}$ drawn i.i.d. from a joint distribution $P_{X \times Y}$ (so that $z_t = (x_t, y_t)$). The marginal distribution P_X is arbitrary and the conditional distribution $P_{Y|X}$ is defined via

$$Y = f^\star(X) + \xi, \tag{2.3}$$

where $f^\star$ belongs to a *model class* $\mathcal{F} \subseteq (\mathcal{X} \to \mathcal{Y})$ and $\mathbb{E}[\xi \mid X] = 0$. The class $\mathcal{F}$ serves as a model for nature. A learning rule takes as input the sample set S and returns a predictor $\widehat{y}_S : \mathcal{X} \to \widehat{\mathcal{Y}}$ (so that $\mathcal{D} = (\mathcal{X} \to \widehat{\mathcal{Y}})$). We define a point-wise loss $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, and the risk of a particular predictor $\widehat{y}$ is given by $\boldsymbol{\ell}(\widehat{y}, z) = \ell(\widehat{y}(x), y)$. The final notion of risk in (2.1) is

$$\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{P} \ell(\widehat{y}_S(x), y)\right].$$

Taking $\widehat{\mathcal{Y}} = \mathcal{Y} = \{\pm 1\}$, $\ell(\widehat{y}, y) = \mathbb{1}\{\widehat{y} \neq y\}$ and $\xi = 0$ recovers PAC learning, while setting $\mathcal{Y} = \mathbb{R}$ and $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ or $\ell(\widehat{y}, y) = |\widehat{y} - y|$ recovers classical nonparametric regression.

The minimax risk is

$$\mathcal{V}_n^{\mathrm{pac}}(\mathcal{F}) = \inf_{\widehat{y}} \sup_{P_X} \sup_{\substack{P_{Y|X} \\ \text{realizable}}} \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{P} \ell(\widehat{y}_S(x), y)\right],$$

while the minimax achievability for a rate $\boldsymbol{\phi}$ is

$$\mathcal{V}_n^{\mathrm{pac}}(\boldsymbol{\phi}) = \inf_{\widehat{y}} \sup_{P_X} \sup_{\substack{P_{Y|X} \\ \text{realizable}}} \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{P} \ell(\widehat{y}_S(x), y) - \boldsymbol{\phi}(x_{1:n}, y_{1:n}, P_X, P_{Y|X})\right]. \tag{2.4}$$

We see that the adaptive rate $\boldsymbol{\phi}$ can depend on the particular draw of examples $\{(x_t, y_t)\}_{t=1}^n$, as well as the true function $f^\star$ and marginal distribution P_X. Nice instances for this setting may include functions $f^\star$ for which the decision boundary is simple—at least simple in areas where the marginal distribution P_X is concentrated (Boucheron et al., 2005)—-or might include instances where the variance of ξ is low.

2.2 Statistical Learning

Existence of a function $f^\star$ that realizes the model $P_{Y|X}$ is a rather strong assumption. We would like to develop methods that enjoy prediction guarantees when this assumption is violated, but hopefully can adapt when the assumption *does* hold.

For classification, the *agnostic PAC framework* (Haussler, 1992; Kearns et al., 1994) generalizes PAC learning to the case where the joint distribution $P_{X\times Y}$ is arbitrary. For regression,

Protocol 1 Statistical Learning

Nature selects distrbution $P_{X\times Y}$.
Learner receives samples $S = (x_1, y_1), \ldots, (x_n, y_n)$ i.i.d. from $P_{X\times Y}$.
Learner returns predictor $\widehat{y}_S \in (\mathcal{X} \to \widehat{\mathcal{Y}})$. (For proper learning, $\widehat{y}_S \in \mathcal{F}$.)

this is referred to as the *misspecified model* setting in nonparametric statistics (Nemirovski, 2000; Tsybakov, 2008) and aggregation (Tsybakov, 2003; Lecué and Rigollet, 2014). The distribution $P_{X\times Y}$ may be completely unrelated to the model class $\mathcal{F}$, and there may indeed be distributions P for which the expected risk $\mathbb{E}\,\ell(f(x), y)$ is large for all $f \in \mathcal{F}$ (for classification, consider the case where labels are drawn uniformly at random). Clasically, instead of looking at minimax risk in the sense of $\boldsymbol{\ell}$, agnostic learning considers *minimax regret*:[1]

$$\mathcal{V}_n^{\text{iid}}(\mathcal{F}) = \inf_{\widehat{y}} \sup_{P_{X\times Y}} \mathbb{E}_S\left[\mathbb{E}\,\ell(\widehat{y}_S(x), y) - \inf_{f\in\mathcal{F}} \mathbb{E}\,\ell(f(x), y)\right]. \tag{2.5}$$

Here the word "regret" reflects that performance is measured relative the class $\mathcal{F}$, which serves as a *benchmark* or *comparator*. A bound on the right-hand-side of this expression is also referred to as an *exact oracle inequality* in statistics (Tsybakov, 2003; Lecué and Rigollet, 2014).

In this case, a natural way to define minimax achievability for an adaptive rate $\mathcal{B}$ is

$$\mathcal{V}_n^{\text{iid}}(\mathcal{F}, \mathcal{B}) = \inf_{\widehat{y}} \sup_{P_{X\times Y}} \mathbb{E}_S\left[\mathbb{E}\,\ell(\widehat{y}_S(x), y) - \inf_{f\in\mathcal{F}} \mathbb{E}\,\ell(f(x), y) - \mathcal{B}(x_{1:n}, y_{1:n}, P_{X\times Y})\right]. \tag{2.6}$$

Note on Terminology. Throughout this thesis we use the symbol ϕ for adaptive rates that bound risk and the symbol $\mathcal{B}$ for adaptive rates that bound regret.

This formulation already subsumes many notions of adaptivity proposed in statistical learning theory (Boucheron et al., 2005). To capture further notions of adaptivity, such as PAC-Bayesian bounds (McAllester, 1999), we can allow the adaptive rate $\mathcal{B}$ to depend on

[1]Note that minimax regret still falls into the statistical decision theory framework for the right choice of L.

the benchmark itself, i.e.

$$\mathcal{V}_n^{\mathrm{iid}}(\mathcal{F}, \mathcal{B}) = \inf_{\widehat{y}} \sup_{P_{X\times Y}} \mathop{\mathbb{E}}_{S} \sup_{f\in\mathcal{F}} [\mathbb{E}\,\ell(\widehat{y}_S(x), y) - \mathbb{E}\,\ell(f(x), y) - \mathcal{B}(f\,; x_{1:n}, y_{1:n}, P_{X\times Y})]. \tag{2.7}$$

Of course, we can go further and return to adaptive rates ϕ that upper bound the expected risk itself, with the understanding that any achievable rate of this form must be large for some instances $P_{X\times Y}$:

$$\mathcal{V}_n^{\mathrm{iid}}(\boldsymbol{\phi}) = \inf_{\widehat{y}} \sup_{P_{X\times Y}} \mathop{\mathbb{E}}_{S} [\mathbb{E}\,\ell(\widehat{y}_S(x), y) - \boldsymbol{\phi}(x_{1:n}, y_{1:n}, P_{X\times Y})]. \tag{2.8}$$

This formulation is syntactically very close to that of $\mathcal{V}_n^{\mathrm{pac}}(\phi)$ in (2.4), but the key difference is that we have dropped the assumption on the conditional distribution.

2.3 Online Supervised Learning

While the agnostic statistical learning setting is certainly more general than the PAC framework, it still makes a strong assumption, namely that the examples $\{(x_t, y_t)\}_{t=1}^n$ are i.i.d. An alternative is *online learning*, where data examples arrive one-by-one and the learner must make predictions on demand, and the data generating process is arbitrary or

Protocol 2 Online Supervised Learning

1: **for** $t = 1, \ldots, n$ **do**
2: Nature provides $x_t \in \mathcal{X}$.
3: Learner selects randomized strategy $q_t \in \Delta(\widehat{\mathcal{Y}})$.
4: Nature provides outcome $y_t \in \mathcal{Y}$.
5: Learner draws $\widehat{y}_t \sim q_t$ and incurs loss $\ell(\widehat{y}_t, y_t)$.
6: **end for**

even adversarial.

The exact setup is as follows. The learner plays n rounds, and for each round t they receive an instance x_t and must produce a prediction $\widehat{y}_t$ using the new instance as well as the previous observations $(x_1, y_1,) \ldots, (x_{t-1}, y_{t-1})$. Nature chooses the true outcome y_t, and the cumulative loss is given by $\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t)$. As in agnostic learning, classical (non-adaptive) online learning evaluates learning procedures based on their *regret* against a benchmark class $\mathcal{F}$:

$$\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t).$$

Like the bit prediction setting in the introduction, we formulate minimax analysis for online learning by imagining a game between the learner and nature. At each round the contribution to regret is a $\max\text{-}\min\text{-}\max$ problem conditioned on the history so far: Nature chooses x_t to maximize regret, then the learner chooses $\widehat{y}_t$ to minimize regret given nature's decision. Finally, nature picks a maximally bad value for y_t based on the learner's decision. The process is repeated for all n rounds, giving rise to the following expression for the minimax value:

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}) = \sup_{x_1}\inf_{\widehat{y}_1}\sup_{y_1}\ldots\sup_{x_n}\inf_{\widehat{y}_n}\sup_{y_n}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t)\right].$$[2]

We adopt the following notation to write expressions of this type more succinctly:

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}) = \left\langle\!\!\left\langle \sup_{x_t}\inf_{\widehat{y}_t}\sup_{y_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t)\right], \tag{2.9}$$

where the notation $\langle\!\langle\star\rangle\!\rangle_{t=1}^{n}$ denotes interleaved application of the operator $\star$ from time $t = 1, \ldots, n$. The difference of the cumulative losses of the forecaster and the loss of any particular benchmark $f \in \mathcal{F}$, which we refer to to as *regret against* f, is denoted $\mathrm{Reg}_n(f) = \sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \ell(f(x_t), y_t)$.

Turning to adaptivity, we can either ask for an adaptive regret bound $\mathcal{B}$ and look at the minimax achievability of $\mathcal{B}$ via

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}, \mathcal{B}) = \left\langle\!\!\left\langle \sup_{x_t}\inf_{\widehat{y}_t}\sup_{y_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t), y_t) - \mathcal{B}(f\,; x_{1:n}, y_{1:n})\right], \tag{2.10}$$

[2]Online supervised learning fits into the decision theory framework by taking $\widehat{\mathcal{Y}} = \mathcal{D}$ and $\boldsymbol{\ell}(\widehat{y}, y) = \ell(\widehat{y}, y)$; note that this learning setting is inherently improper.

or we can ask for a general adaptive risk bound ϕ and look at minimax achievability via

$$\mathcal{V}_n^{\mathrm{ol}}(\phi) = \left\langle\!\!\left\langle \sup_{x_t} \inf_{\widehat{y}_t} \sup_{y_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{1}{n} \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_{1:n}, y_{1:n}) \right]. \tag{2.11}$$

From this definition, we see that for any ϕ there always exists an algorithm that guarantees

$$\frac{1}{n} \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) \le \phi(x_{1:n}, y_{1:n}) + \mathcal{V}(\phi) \quad \forall \text{ sequences } x_{1:n}, y_{1:n}.$$

In a slightly more general setting, Protocol 2, we allow the learner to be randomized, i.e. to select a distribution q_t from which the prediction $\widehat{y}_t$ is sampled only *after* y_t.[3] For such randomized learners, minimax achievability is written

$$\mathcal{V}_n^{\mathrm{ol}}(\phi) = \left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t} \sup_{y_t} \mathop{\mathbb{E}}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{1}{n} \sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_{1:n}, y_{1:n}) \right]. \tag{2.12}$$

The online-to-batch principle, dating back to the very genesis of learning theory (Vapnik and Chervonenkis, 1968), implies that for any rate $\phi(x_{1:n}, y_{1:n})$,

$$\mathcal{V}_n^{\mathrm{pac}}(\phi) \le \mathcal{V}_n^{\mathrm{iid}}(\phi) \le \mathcal{V}_n^{\mathrm{ol}}(\phi). \tag{2.13}$$

An important development in this thesis is that while there are indeed rates ϕ for which $\mathcal{V}_n^{\mathrm{ol}}(\phi) \gg \mathcal{V}_n^{\mathrm{iid}}(\phi)$, for many types of adaptivity of practical interest we have $\mathcal{V}_n^{\mathrm{ol}}(\phi) \le c \cdot \mathcal{V}_n^{\mathrm{iid}}(\phi)$ for some small constant c, meaning that even when offline learning is the end goal we pay essentially no price for considering the more general framework, and can therefore leverage the advantages (e.g. single pass learning) that it provides.

2.4 Online Convex Optimization

Online convex optimization (OCO) is close relative of the online supervised learning setting, in which a learner makes vector-valued predictions and is evaluated against

[3]This setting is necessary to handle various technical issues such as non-convex losses.

an adversarially chosen sequence of convex loss functions. This model is useful for solving large-scale empirical risk minimization problems for machine learning, as well as for directly performing minimization of the population risk in statistical learning and stochastic optimization. In particular, regret bounds in the online convex optimization immediately imply upper bounds on the *oracle complexity* of stochastic convex optimization.

We describe a randomized variant of the OCO setting here. We play n rounds $t = 1, \dots, n$. At each round the learner chooses a distribution q_t over a convex set $\mathcal{W}$. Nature chooses a convex loss f_t, and the learner samples $w_t \sim q_t$ and experiences loss $f_t(w_t)$. Depending on the application nature may be constrained to choose, for example, 1-Lipschitz or 1-smooth convex functions. We let $\mathcal{Z}$ denote their set of decisions.

Protocol 3 Online Convex Optimization

for $t = 1, \dots, n$ **do**
  Learner selects strategy $q_t \in \Delta(\mathcal{W})$ for convex decision set $\mathcal{W}$.
  Nature selects convex loss $f_t \colon \mathcal{W} \to \mathbb{R}$.
  Learner draws $w_t \sim q_t$ and incurs loss $f_t(w_t)$.
end for

In online convex optimization the usual notion of performance is regret relative to the benchmark constraint set $\mathcal{W}$. In particular, the (non-adaptive) minimax value is given by

$$\mathcal{V}_n^{\mathrm{oco}}(\mathcal{W}) = \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{W})} \sup_{f_t \in \mathcal{Z}} \mathop{\mathbb{E}}_{w_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{1}{n} \sum_{t=1}^{n} f_t(w_t) - \inf_{w \in \mathcal{W}} \frac{1}{n} \sum_{t=1}^{n} f_t(w) \right]. \tag{2.14}$$

To analyze adaptive regret bounds $\mathcal{B}$, minimax achievability is defined through

$$\mathcal{V}_n^{\mathrm{oco}}(\mathcal{W}, \mathcal{B}) = \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{W})} \sup_{f_t \in \mathcal{Z}} \mathop{\mathbb{E}}_{w_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{w \in \mathcal{W}} \left[\frac{1}{n} \sum_{t=1}^{n} f_t(w_t) - \frac{1}{n} \sum_{t=1}^{n} f_t(w) - \mathcal{B}(w\,;\, f_{1:n}) \right], \tag{2.15}$$

and for adaptive risk bounds minimax achievability is given by

$$\mathcal{V}_n^{\mathrm{oco}}(\boldsymbol{\phi}) = \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{W})} \sup_{f_t \in \mathcal{Z}} \mathop{\mathbb{E}}_{w_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\frac{1}{n} \sum_{t=1}^{n} f_t(w_t) - \boldsymbol{\phi}(f_{1:n}) \right]. \tag{2.16}$$

Online convex optimization can be thought of as a special case of the online supervised learning setting when there are no contexts ($\mathcal{X} = \{\emptyset\}$), but the settings are distinguished in literature by a focus on appropriately handling the complexity of the output space, which is typically high-dimensional in OCO applications.

2.5 Contextual Bandits

Online supervised learning and online convex optimization are very general and powerful models that are useful both for streaming learning settings and (via online-to-batch) offline statistical learning. One drawback is that both settings make the assumption that the entire loss *function* $\boldsymbol{\ell}(\cdot, z_t)$ is observable, while in many applications we may only observe the *value* $\boldsymbol{\ell}(\widehat{y}_t, z_t)$ under the learner's decision. Such a model is appropriate in news article recommendation and related sequential decision making problems: The learner repeatedly suggests news articles to users on a website and would like to improve their performance over time, but they only observe whether each user views the article that was suggested, not which of the potential articles the user would have preferred in hindsight (Li et al., 2010). The *contextual bandit* model formalizes this problem.

Protocol 4 Contextual Bandit

for $t = 1, \ldots, n$ **do**
    Nature provides context $x_t \in \mathcal{X}$.
    Learner selects randomized strategy $q_t \in \Delta(\mathcal{A})$.
    Nature provides outcome $\ell_t \in \mathcal{L} \subset \mathbb{R}_+^{\mathcal{A}}$.
    Learner draws action $a_t \sim q_t$ and observes loss $\ell_t(a_t)$.
end for

To describe the setting we adopt standard notation from contextual bandit literature (e.g. Agarwal et al. (2014)). The learner plays n rounds. In each round t they receive an instance x_t and must select a discrete action $a_t \in \mathcal{A} := [K]$ using the new instance as well as the previous observations. We allow the learner to randomize and denote their distribution $q_t \in \Delta(\mathcal{A})$. Once the distribution is selected, nature chooses a non-negative loss vector

$\ell_t \in \mathcal{L} \subset \mathbb{R}_+^{\mathcal{A}}$, and then the learner samples $a_t \sim q_t$. Their instantaneous loss is $\ell_t(a_t)$, and the do not observe $\ell_t(a)$ for actions $a \neq a_t$.

Contextual bandit literature focuses on regret against a benchmark class of discrete policies $\Pi \subseteq (\mathcal{X} \to \mathcal{A})$. In particular minimax regret is defined via

$$\mathcal{V}_n^{\mathrm{cb}}(\Pi) = \inf_{\boldsymbol{q}} \sup_{\overline{\boldsymbol{\ell}}} \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell_t(a_t) - \inf_{\pi\in\Pi} \frac{1}{n}\sum_{t=1}^{n} \ell_t(\pi(x_t))\right],$$

where the infimum ranges over all randomized learner policies $\boldsymbol{q}$ (i.e., sequences of mappings from past outcomes to action distributions), the supremum ranges over all adversary policies $\overline{\boldsymbol{\ell}}$, and the expectation is with respect to the learner's randomization.[4] We define minimax achievability for adaptive regret bounds $\mathcal{B}$ and adaptive risk bounds $\boldsymbol{\phi}$ via

$$\mathcal{V}_n^{\mathrm{cb}}(\Pi, \mathcal{B}) = \inf_{\boldsymbol{q}} \sup_{\overline{\boldsymbol{\ell}}} \mathbb{E} \sup_{\pi\in\Pi}\left[\frac{1}{n}\sum_{t=1}^{n} \ell_t(a_t) - \frac{1}{n}\sum_{t=1}^{n} \ell_t(\pi(x_t)) - \mathcal{B}(\pi\,;\ell_{1:n})\right],$$

and

$$\mathcal{V}_n^{\mathrm{cb}}(\boldsymbol{\phi}) = \inf_{\boldsymbol{q}} \sup_{\overline{\boldsymbol{\ell}}} \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell_t(a_t) - \boldsymbol{\phi}(\ell_{1:n})\right],$$

respectively.

2.6 The Minimax Theorem

Defining minimax achievability is a useful first step, but how can we actually derive bounds on this value, e.g. on $\mathcal{V}_n^{\mathrm{ol}}(\boldsymbol{\phi})$? The starting point of the approach we develop in Part II is the *minimax theorem*, which allows us to exchange the order of the $\min$ and $\max$ player in repeated $\min$-$\max$ expressions such as (2.11). The version of the theorem we use is due to Sion (Sion, 1958), and generalizes the classical minimax theorem of Von Neumann.

[4] The partial feedback in contextual bandits prevents one from writing the minimax value in a step-by-step fashion along the lines of (2.9), which is why we adopt the policy formulation here.

Theorem 1 (Sion's Minimax Theorem). *Let $F : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, where $\mathcal{U}$ is a convex and compact subset of a linear topological space and and $\mathcal{V}$ is convex subset of a linear topological space. If $F(u, \cdot)$ is upper semicontinuous and quasiconcave over $\mathcal{V}$ for all $u \in \mathcal{U}$ and $F(\cdot, v)$ is lower semicontinuous and quasiconvex over $\mathcal{U}$ for all $v \in \mathcal{V}$, then*

$$\min_{u \in \mathcal{U}} \sup_{v \in \mathcal{V}} F(u, v) = \sup_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} F(u, v). \tag{2.17}$$

Our use of the minimax theorem throughout this thesis follows an approach pioneered by Abernethy et al. (2008) and Rakhlin et al. (2010). We repeatedly invoke the following fact in our results for the online learning and online convex optimization models.

Proposition 1. Let $F : (\mathcal{X} \times \mathcal{Y} \times \widehat{\mathcal{Y}})^n \to \mathbb{R}$ be a uniformly bounded function. Let $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ be compact. Then

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta_{\widehat{\mathcal{Y}}}} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} F((x_1, y_1, \widehat{y}_1), \ldots, (x_n, y_n, \widehat{y}_n)) \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_{\mathcal{Y}}} \inf_{\widehat{y}_t \in \widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} F((x_1, y_1, \widehat{y}_1), \ldots, (x_n, y_n, \widehat{y}_n)).
\end{aligned}$$

2.7 Chapter Notes

There are many further learning models that will not be covered in detail in this thesis, and for which extending the techniques we present is an interesting direction for future research. Notable examples include various combinations of the settings discussed in this chapter, including stochastic contextual bandits and stochastic optimization, bandit linear optimization, and bandit convex optimization, as well as other interactive learning settings such as active learning (Hanneke, 2014; Krishnamurthy et al., 2017) and reinforcement learning (Szepesvári, 2010). It is straightforward to extend the definition of minimax achievability to these settings and beyond.

Part II

Equivalence of Prediction, Martingales, and Geometry

CHAPTER 3

OVERVIEW OF PART II

In this part of the thesis we introduce a new equivalence between adaptive online learning, martingale inequalities, and Burkholder functions (recall Figure 1.1). This allows to systematically:

1. Characterize minimax achievability of adaptive learning guarantees in learning models of practical importance, and do so in an algorithm-independent fashion.
2. Develop efficient algorithms to obtain achievable adaptive rates.

The roadmap for the development of the equivalence is as follows.

- First, in Chapter 4 we prove the equivalence for a simplified version of the online supervised learning setting.
- In Chapter 5 we extend the equivalence to the general online supervised learning setting, and also introduce a notion of *sufficient statistics* for online learning. The development of sufficient statistics allows us to deduce additional geometric properties for Burkholder functions when we apply the Burkholder method, and as a consequence leads to online learning algorithms with reduced memory requirements.
- Finally, in Chapter 6 we develop probabilistic tools to directly prove upper bounds on martingale inequalities that arise in the equivalence framework. These tools can be used to certify achievability (and thus existence of Burkholder functions) for concrete adaptive rates of interest without necessarily exhibiting an explicit algorithm, and in particular allow us to deduce achievability of adaptive rates for complicated models (e.g., decision trees and neural networks) where it is not possible to derive

computationally efficient algorithms with worst-case performance guarantees. The probabilistic tools of Chapter 6 play a role similar to that of empirical process theory the classical statistical learning model.

While the tools in this part of the thesis are specific to online learning, they form the backbone for all of the new adaptive learning guarantees and algorithms we develop throughout Part III. Indeed, we use the tools to derive powerful consequences for tasks beyond online learning, including statistical learning, optimization, boosting, and bandits. Focusing on the online learning setting for now is advantageous for the following reasons.

- Considering a more powerful adversary than in e.g., the batch statistical learning setting, turns out to lead to stronger analysis tools. In particular, the equivalence we develop through the Burkholder method exploits that the learner must adapt to data generated by an arbitrary and potentially adaptive adversary.
- Online learning can be used to solve both offline statistical learning (via online-to-batch conversion (Vapnik and Chervonenkis, 1968)) and contextual bandits (via importance-weighting reductions (Auer et al., 2002b)). Even when our goal is to solve the batch statistical learning setting (i.e., to give bounds on $\mathcal{V}_n^{\mathrm{iid}}(\phi)$), we do not lose much by focusing on the harder online learning setting because we derive *adaptive rates*. The adaptive matrix prediction example we develop in this chapter exemplifies this phenomenon.

CHAPTER 4

THE EQUIVALENCE

This chapter develops the equivalence of adaptive learning, martingale inequalities, and Burkholder functions in a simplified version of the general online supervised learning setting. The purpose of this simplification is to communicate the key ideas and techniques while keeping exposition as simple as possible. Using the *online collaborative filtering* task as a running example, we first propose a new notion of adaptivity, then derive an efficient algorithm to achieve this new type of adaptivity using the Burkholder method.

We consider the following setup, which is a special case of Protocol 2 where we restrict to *linear losses*. At each time $t = 1, \dots, n$ the learner receives $x_t \in \mathcal{X}$, which we take to be a subset of some vector space. The learner predicts $\widehat{y}_t \in \mathbb{R}$, and receives an outcome $y_t \in \{\pm 1\}$. Performance is measured via the linear loss $\ell(\widehat{y}_t, y_t) = -\widehat{y}_t \cdot y_t$, and the goal is a adaptive prediction guarantee of the form[1]

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t \leq \boldsymbol{\phi}(y_1 x_1, \dots, y_n x_n) \qquad \text{for all sequences } x_{1:n}, y_{1:n}. \tag{4.1}$$

Note that the adaptive rate $\boldsymbol{\phi}$ depends on x_t and y_t through the product $y_t x_t$. This structure simplifies presentation, but will be relaxed in Chapter 5.

4.1 Running Example: Matrix Prediction

To ground our development in the reality, let us show how the online learning setting in (4.1) can be used to solve an online version of the classical *collaborative filtering* problem (Billsus and Pazzani, 1998). The online collaborative filtering problem proceeds as follows.

[1] There is no need to allow for randomized learners for the development in this chapter. This extra level of generality is covered in Chapter 5.

At each time $t = 1, \ldots, n$ we receive a user-movie pair $(i_t, j_t) \in [d_1] \times [d_2]$. Our goal is to predict the user's affinity for the movie. We predict a value $\widehat{y}_t \in \mathbb{R}$ and receive an outcome $y_t \in \{\pm 1\}$, with a value of $+1$ indicating that the user likes the movie and -1 indicating that they dislike it. The instantaneous loss is $\ell(\widehat{y}_t, y_t) = -\widehat{y}_t \cdot y_t$, so it is in our best interest to confidently predict the same sign as y_t.

To develop algorithms, we adopt the well-known approach of formulating collaborative filtering as matrix factorization (Azar et al., 2001; Rennie and Srebro, 2005; Srebro and Shraibman, 2005; Hazan et al., 2012). A ubiquitous assumption in these works and beyond is that the underlying matrix of user-movie affinities is modeled well by a low-rank matrix. We do not explicitly make such an assumption, but we will develop algorithms that predict well whenever this is the case. Specifically, letting $x_t = e_{i_t} e_{j_t}^\top \in \mathbb{R}^{d_1 \times d_2}$ denote the incidence matrix for the user-movie pair at time t, we focus on adaptive regret guarantees of the form

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t \leq \inf_{w : \|w\|_\Sigma \leq \tau} \sum_{t=1}^{n} -\langle w, x_t \rangle \cdot y_t + \mathcal{B}(y_1 x_1, \ldots, y_n x_n), \tag{4.2}$$

where $\|\cdot\|_\Sigma$ denotes the nuclear norm, $\langle w, x \rangle$ is the standard matrix inner product, and $\mathcal{B}$ is an adaptive regret bound. In particular, to predict as well as the best rank-r matrix with entry magnitudes bounded by 1 (up to the regret bound $\mathcal{B}$), it suffices to take $\tau = \sqrt{r d_1 d_2}$ in this expression. Note that we have chosen to compete with the set of all nuclear norm bounded matrices rather than explicitly competing with the set of all low-rank matrices because this typically leads to computationally efficient algorithms; this is the standard approach in high-dimensional statistics (Candès and Recht, 2009; Candes and Plan, 2010; Foygel and Srebro, 2011).

For *linear prediction problems* with the structure in (4.2), the algorithmic workhorses in online learning are the *mirror descent* and *dual averaging/follow-the-regularized-leader* families of algorithms (Nemirovski et al., 1983; Ben-Tal and Nemirovski, 2001; Hazan, 2016). To illustrate the necessity of new algorithmic ideas for adaptive learning, let us examine how mirror descent fares for the matrix prediction problem. Using standard techniques (Arora et al., 2012; Hazan et al., 2012) it follows that mirror descent with the matrix entropy regularizer, also known as the *matrix multiplicative weights* strategy, can ensure an inequality of the form (4.2) with

$$\mathcal{B}(y_1x_1, \ldots, y_nx_n) \propto \tau \cdot \sqrt{\log(d_1 + d_2) \cdot \sum_{t=1}^{n} \|y_tx_t\|_\sigma^2}. \tag{4.3}$$

For our setting, $\|y_tx_t\|_\sigma = 1$ for all t, and so we have $\mathcal{B}(x_{1:n}) \approx \sqrt{nrd_1d_2}$. This is somewhat unfortunate: With such an bound our average regret will not drop below one until we have seen every entry in the matrix! This is not the end of the story, however. It turns out that mirror descent is missing part of the problem geometry. Let us investigate further.

- On one hand, it turns out that the rate $\sqrt{nrd_1d_2}$ is minimax optimal, meaning there are indeed sequences for which any algorithm competing with the nuclear norm ball must have poor regret. In other words, any achievable function $\mathcal{B}(y_1x_1, \ldots, y_nx_n)$ must be large for some sequences.
- On the other hand, it is known that in the batch statistical learning setting, if the observed entries are generated uniformly at random i.i.d., it is possible to give a bound of the form (4.2) with $\mathcal{B} \approx \sqrt{nr\max\{d_1, d_2\}}$. This is a significant improvement, and guarantees that we generalize after seeing roughly $r \cdot \max\{d_1, d_2\}$ entries.

We will develop an algorithm that smoothly interpolates between the "nice data" regime above (where entries are uniform and i.i.d.) and the minimax rate (4.3), without having to know in advance whether data is nice.

The algorithm is developed by leveraging the equivalence of prediction inequalities, martingale inequalities, and Burkholder functions. In particular, the equivalence will reveal that adaptive algorithms that exploit the niceness described above are closely related to the so-called matrix Khintchine inequalities (Lust-Piquard and Pisier, 1991; Tropp, 2012; Mackey et al., 2014).

4.2 Emergence of Martingales

As a first step toward proving the equivalence, we show that for any adaptive rate $\boldsymbol{\phi}$, achievability of a prediction guarantee of the form (4.1) implies a certain algorithm-independent inequality involving $\boldsymbol{\phi}$ that we call a *generalized martingale inequality*.

Proposition 2. Let an adaptive rate $\boldsymbol{\phi}$ be fixed, and suppose that for any $n \geq 1$ there exists some algorithm that attains the prediction inequality (4.1). Then it holds that

$$\inf_{n} \inf_{\boldsymbol{x}_{1:n}} \mathbb{E}_{\epsilon}[\boldsymbol{\phi}(\epsilon_1 \boldsymbol{x}_1(\epsilon), \ldots, \epsilon_n \boldsymbol{x}_n(\epsilon))] \geq 0. \tag{4.4}$$

Here the infimum ranges over all $\mathcal{X}$-valued *predictable processes* or *trees* $\boldsymbol{x}$ of the form $(\boldsymbol{x}_t)_{t=1}^{n}$, where $\boldsymbol{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$, and $\epsilon \in \{\pm 1\}^n$ is a sequence of Rademacher random variables.

Note on Terminology. For the remainder of this thesis, we adopt the shorthand $\boldsymbol{x}_t(\epsilon) := \boldsymbol{x}_t(\epsilon_{1:t-1})$ for predictable processes $\boldsymbol{x}$. When the dependence is clear from context, we abbreviate further to $\boldsymbol{x}_t := \boldsymbol{x}_t(\epsilon)$. We call (4.4) a generalized martingale inequality because the sequence $(\epsilon_t \boldsymbol{x}_t)_{t=1}^{n}$ is a martingale difference sequence. Martingales with this structure are sometimes called *dyadic martingales* or *Paley-Walsh martingales* (Hytönen et al., 2016).

Proof. The proof is quite simple, and follows the same reasoning used for Cover's characterization in Section 1.5. The idea is that since (4.1) is guaranteed to hold for every sequence $x_{1:n}$, $y_{1:n}$, it must hold in particular for a class of sequences that we are free to choose.

We first draw a sequence of independent Rademacher random variables $\epsilon \in \{\pm 1\}^n$ and set $y_t = \epsilon_t$. We then pick an arbitrary $\mathcal{X}$-valued predictable process $\boldsymbol{x}$ and set $x_t = \boldsymbol{x}_t(\epsilon)$. With this choice, (4.1) implies that for every draw of ϵ,

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot \epsilon_t \leq \boldsymbol{\phi}(\epsilon_1 \boldsymbol{x}_1, \ldots, \epsilon_n \boldsymbol{x}_n).$$

In the supervised learning model the prediction $\widehat{y}_t$ can only depend on $\epsilon_{1:t-1}$. Consequently, the left-hand-side of this inequality has mean zero for *any* algorithm, and so taking expectation over the draw of ϵ leads us to conclude

$$\mathop{\mathbb{E}}_{\epsilon}[\boldsymbol{\phi}(\epsilon_1 \boldsymbol{x}_1, \ldots, \epsilon_n \boldsymbol{x}_n)] \geq 0.$$

The result (4.4) follows because the choice of $\boldsymbol{x}$ in this argument is arbitrary. $\square$

In Section 4.5 we show that (4.4) is also *necessary*. This gives us a powerful modeling tool: To check achievability of the rate $\boldsymbol{\phi}$, it suffices to check that algorithm-independent inequality (4.4) holds, and conversely if the inequality does not hold the rate is not achievable. Before proving this fact, we spend a moment developing generalized martingale inequalities—of which (4.4) is a special case—in more detail.

4.3 Generalized Martingale Inequalities

Let $(X_t)_{t\geq 1}$ be a martingale difference sequence; that is, a sequence of vector-valued random variables with the property that $\mathbb{E}[X_t \mid X_1, \ldots, X_{t-1}] = 0$ for all $t \geq 1$. We define a *generalized martingale inequality* to be any inequality of the form

$$\mathbb{E}\, V(X_1, \ldots, X_n) \leq 0 \quad \forall n \geq 1, \tag{4.5}$$

where $V : \cup_{n\geq 1}\mathcal{X}^n \to \mathbb{R}$ is some fixed function. Observe that the necessary condition (4.4) falls into this format by selecting $V = -\boldsymbol{\phi}$. In fact, many familiar inequalities are captured by (4.5).

Example 1 (Azuma-Hoeffding Inequality)**.** *The Azuma-Hoeffding inequality (e.g. (Boucheron et al., 2013)) can be written as the observation that*

$$\mathbb{E}\exp\left(\sum_{t=1}^{n}\epsilon_t\boldsymbol{x}_t - \frac{\boldsymbol{x}_t^2}{2}\right) \leq 1.$$

for all $n \geq 1$ and all real-valued predictable processes $\boldsymbol{x}$. The corresponding function V for this inequality is

$$V(x_1, \ldots, x_n) = \exp\left(\sum_{t=1}^{n} x_t - \frac{x_t^2}{2}\right) - 1.$$

Example 2 (Nemirovski's Inequality)**.** *Nemirovski's inequality (Nemirovski, 2000; Boucheron et al., 2013) states that if $\|\cdot\|$ is any smooth norm*[2] *in a Banach space $\mathfrak{B}$, there exists a constant $C_{\mathfrak{B}}$ such that*

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|^2 \leq C_{\mathfrak{B}} \cdot \sum_{t=1}^{n}\mathbb{E}\|X_t\|^2. \tag{4.6}$$

for all $\mathfrak{B}$-valued martingale difference sequences $(X_t)_{t=1}^n$ and all $n \geq 1$. The corresponding function V is

$$V(x_1, \ldots, x_n) = \left\|\sum_{t=1}^{n} x_t\right\|^2 - C_{\mathfrak{B}} \cdot \sum_{t=1}^{n}\|x_t\|^2.$$

4.4 The Burkholder Method

Burkholder's method gives a characterization of the generalized martingales (4.5) developed in the previous section. Informally, the characterization states that whenever such an inequality holds for a particular function V, there exist certain "extremal functions" that a) imply a strengthened version of the original inequality and b) enjoy a certain

[2] Cf. Section 1.7.

"restricted concavity" property. In a little more detail, Burkholder's characterization has two implications:

1. Given a function V, we can check if it has the restricted concavity property. If not, there exists a *Burkholder function* $\mathbf{U}$ that *does* have this property, and for which $V \le \mathbf{U}$ and $\mathbb{E}\,\mathbf{U} \le 0$.
2. Given a function $\mathbf{U}$ with the desired concavity property, there is a simple inductive proof of the inequality $\mathbb{E}\,V \le 0$.

The precise characterization—specialized to dyadic martingales for ease of presentation—is as follows.

Theorem 2 (Burkholder's Characterization (Burkholder, 1981, 1984, 1986, 1991))**.** *The following statements are equivalent:*

- *The inequality*

$$\mathbb{E}\,V(\epsilon_1 \boldsymbol{x}_1, \ldots, \epsilon_n \boldsymbol{x}_n) \le 0 \tag{4.7}$$

holds for all predictable processes $\boldsymbol{x}$ and $n \ge 1$.

- *There exists a function $\mathbf{U} : \cup_{n \ge 0} \mathcal{X}^n \to \mathbb{R}$ such that*

  1^o $V(x_1, \ldots, x_n) \le \mathbf{U}(x_1, \ldots, x_n)$ *for all* $x_1, \ldots, x_n$ *and* $n \ge 1$.

  2^o $\mathbf{U}(\emptyset) \le 0$.

  3^o *For all* $x_1, \ldots, x_n$*,* $n \ge 0$*, and* $x \in \mathcal{X}$*,*

$$\mathbb{E}\,\mathbf{U}(x_1, \ldots, x_n, \epsilon x) \le \mathbf{U}(x_1, \ldots, x_n). \qquad \textit{(restricted concavity)}$$

Example 3 (Example 1 and Example 2 continued)**.** *In fact, for both Example 1 and Example 2 the functions V already satisfy properties 1^o/2^o/3^o, and therefore cannot be strengthened.*

Properties $1^o$and $2^o$are immediate for both functions. For Example 1 property 3^o follows using the standard moment generating function bound $\mathbb{E}_\epsilon e^x \leq e^{\frac{x^2}{2}}$. For Example 2, 3^o follows by using the geometric propertythat for any β-smooth norm, the function $\Psi(x) = \frac{1}{2}\|x\|^2$ satisfies $\Psi(x) \leq \Psi(y) + \langle \nabla\psi(x), y - x\rangle + \frac{\beta}{2}\|x\|^2$.

We now prove Theorem 2.

Proof. *Burkholder Function $\implies$ Martingale Inequality.*

Let the process $\boldsymbol{x}$ and n be fixed. As promised, the properties of the Burkholder function lend themselves to a simple proof of the martingale inequality via step-by-step peeling.

$$\mathbb{E}_\epsilon V(\epsilon_1\boldsymbol{x}_1, \ldots, \epsilon_n\boldsymbol{x}_n) \overset{1^\circ}{\leq} \mathbb{E}_\epsilon \mathbf{U}(\epsilon_1\boldsymbol{x}_1, \ldots, \epsilon_n\boldsymbol{x}_n) \overset{3^\circ}{\leq} \mathbb{E}_\epsilon \mathbf{U}(\epsilon_1\boldsymbol{x}_1, \ldots, \epsilon_{n-1}\boldsymbol{x}_{n-1}) \overset{3^\circ}{\leq} \ldots \overset{3^\circ}{\leq} \mathbf{U}(\emptyset) \overset{2^\circ}{\leq} 0.$$

Martingale Inequality $\implies$ Burkholder Function.

We exhibit a particular choice for the Burkholder function which will be shown to satisfy the desired properties:

$$\mathbf{U}^\star(x_1, \ldots, x_t) = \sup_{n\geq t} \sup_{\boldsymbol{x}_{t+1:n}} \mathop{\mathbb{E}}_{\epsilon_{t+1:n}} V(x_1, \ldots, x_t, \epsilon_{t+1}\boldsymbol{x}_{t+1}, \ldots, \epsilon_n\boldsymbol{x}_n).$$

Property 1^o is immediately implied by this definition, since we can set $n = t$. Property 2^o follows because the inequality (4.7) is assumed to hold, and because $\mathbf{U}^\star(\emptyset) = \sup_{n\geq 1}\sup_{\boldsymbol{x}_{1:n}} \mathbb{E}_{\epsilon_{1:n}} V(\epsilon_1\boldsymbol{x}_1, \ldots, \epsilon_n\boldsymbol{x}_n)$. Property 3^o is where the definition of $\mathbf{U}^\star$ helps out the most. For any $x_1, \ldots, x_{t-1}$, and $x \in \mathcal{X}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\epsilon} \mathbf{U}^\star(x_1, \ldots, x_{t-1}, \epsilon x) &= \mathop{\mathbb{E}}_{\epsilon} \sup_{n\geq t} \sup_{\boldsymbol{x}_{t+1:n}} \mathop{\mathbb{E}}_{\epsilon_{t+1:n}} V(x_1, \ldots, \epsilon x, \epsilon_{t+1}\boldsymbol{x}_{t+1}, \ldots, \epsilon_n\boldsymbol{x}_n) \\
&\leq \sup_{n\geq t-1} \sup_{\boldsymbol{x}_{t:n}} \mathop{\mathbb{E}}_{\epsilon_{t:n}} V(x_1, \ldots, \epsilon\boldsymbol{x}_t, \epsilon_{t+1}\boldsymbol{x}_{t+1}, \ldots, \epsilon_n\boldsymbol{x}_n) \\
&= \mathbf{U}^\star(x_1, \ldots, x_{t-1}).
\end{aligned}$$

The inequality here holds because we can fold the expectation of the increment ϵx into the supremum in the definition $\mathbf{U}^\star$.

$\square$

Note that the proof of Theorem 2 in fact provides a construction for the "optimal" function $\mathbf{U}$, but it is not clear how to directly evaluate the optimal function efficiently (see Section 5.8 for a discussion of the computational prospects of automating this process).

Interestingly, an result by Pisier predating that of Burkholder can be seen as an application of Burkholder's method to the special case of Nemirovski-style inequalities (which belong to the class of *martingale type* inequalities in Banach space literature) (Pisier, 1975). Pisier showed that for any norm $\|\cdot\|$, not a-priori known to be smooth, that if the inequality (4.6) holds then there exists a smooth function that majorizes $\frac{1}{2}\|x\|^2$ and satisfies the other properties of Theorem 2.

While the examples in this section are quite simple, in a moment we will show that for the matrix prediction setup, the natural function V does not have the desired concavity property, and therefore Burkholder's method yields a strengthened inequality. First, we show how the Burkholder function $\mathbf{U}$ can be used for prediction.

4.5 The Burkholder Algorithm

We have already shown that achievability of the adaptive rate ϕ implies a generalized martingale inequality (4.1). We now close the loop and show that the generalized martingale inequality is also *sufficient* for achievability. This theorem is algorithmic in nature: It yields a strategy that attains the adaptive rate ϕ, and that is efficient in terms of queries to $\mathbf{U}$. We call the resulting strategy *the Burkholder algorithm*.

Theorem 3 (Burkholder Algorithm)**.** *Let an adaptive rate ϕ be fixed. Then the following* *statements are equivalent:*

- *For every $n \geq 1$ there exists some algorithm that attains the prediction inequality (4.1).*
- *The following generalized martingale inequality holds:*

$$\inf_{n} \inf_{\boldsymbol{x}_{1:n}} \mathbb{E}_{\epsilon}[\boldsymbol{\phi}(\epsilon_1 \boldsymbol{x}_1, \ldots, \epsilon_n \boldsymbol{x}_n)] \geq 0.$$

Furthermore, the following strategy achieves $\boldsymbol{\phi}$ whenever it is achievable:

1. *Find a Burkholder function $\mathbf{U}$ for $V := -\boldsymbol{\phi}$.*
2. *At each time t, play*

$$\widehat{y}_t = \frac{\mathbf{U}(y_1 x_1, \ldots, -x_t) - \mathbf{U}(y_1 x_1, \ldots, +x_t)}{2}.$$

Proof. We prove that the martingale inequality implies achievability of $\boldsymbol{\phi}$. First, we invoke Theorem 2, which implies that there exists a Burkholder function $\mathbf{U}$ for $V = -\boldsymbol{\phi}$. The remainder of the proof is to show that the Burkholder algorithm, run with $\mathbf{U}$, achieves $\boldsymbol{\phi}$. To begin, Property 1^o clearly implies

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t - \boldsymbol{\phi}(y_1 x_1, \ldots, y_n x_n) \leq \sum_{t=1}^{n} -\widehat{y}_t \cdot y_t + \mathbf{U}(y_1 x_1, \ldots, y_n x_n).$$

We now analyze the contribution of the final round n to the gap above. Let x_n be fixed. Then clearly the best strategy $\widehat{y}$ given the history so far is to solve

$$\min_{\widehat{y} \in \mathbb{R}} \max_{y_n \in [-1,+1]} [-\widehat{y}_n \cdot y_n + \mathbf{U}(y_1 x_1, \ldots, y_n x_n)],$$

which we rewrite as

$$\min_{\widehat{y} \in \mathbb{R}} \max\{-\widehat{y}_n + \mathbf{U}(y_1 x_1, \ldots, +x_n), \widehat{y}_n + \mathbf{U}(y_1 x_1, \ldots, -x_n)\}.$$

The solution is to set the two terms inside the max equal, which leads to

$$\widehat{y}_n = \frac{\mathbf{U}(y_1 x_1, \ldots, -x_n) - \mathbf{U}(y_1 x_1, \ldots, +x_n)}{2}.$$

Plugging in this strategy gives the inequality

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t + \mathbf{U}(y_1 x_1, \ldots, y_n x_n) \leq \sum_{t=1}^{n-1} -\widehat{y}_t \cdot y_t + \mathop{\mathbb{E}}_{\epsilon} \mathbf{U}(y_1 x_1, \ldots, \epsilon x_n).$$

To proceed, we use Property 3^o to arrive at any upper bound of

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t + \mathbf{U}(y_1 x_1, \ldots, y_n x_n) \leq \sum_{t=1}^{n-1} -\widehat{y}_t \cdot y_t + \mathop{\mathbb{E}}_{\epsilon} \mathbf{U}(y_1 x_1, \ldots, x_{n-1}).$$

Repeating this argument back until time $t = 1$, we get

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t + \mathbf{U}(y_1 x_1, \ldots, y_n x_n) \leq \mathbf{U}(\emptyset) \overset{2^\circ}{\leq} 0,$$

and so the rate is achieved.

□

In addition to being simple and elegant, this algorithm enjoys the additional property of being horizon-independent. We also remark that the argument suggests a set of adaptive rates that are *pareto-optimal* within the class of all achievable rates. Namely, if ϕ is achievable, the following strictly stronger inequality is also achievable:

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t \leq \; -\mathbf{U}(y_1 x_1, \ldots, y_n x_n) \qquad \text{for all sequences } x_{1:n}, y_{1:n}, \tag{4.8}$$

where $\mathbf{U}$ is any Burkholder function for $V = -\phi$.

In fact, it turns out that we can directly prove that (4.1) is sufficient for achievability of the rate ϕ without directly invoking the Burkholder method. This is explored in the next chapter.

4.6 Burkholder Function for Matrix Prediction

We now return to the matrix prediction setting in (4.2). We derive a new efficient and adaptive algorithm by exhibiting an explicit Burkholder function for the problem. To proceed we must first nail down our choice of adaptive rate. We do so with the help of the equivalence. Recall that our desired prediction guarantee has the form

$$\phi(y_1x_1,\ldots,y_nx_n) = \inf_{w:\|w\|_\Sigma\le\tau}\sum_{t=1}^{n} -\langle w,x_t\rangle\cdot y_t + \mathcal{B}(y_1x_1,\ldots,y_nx_n),$$

where $\mathcal{B}$ has yet to be decided. We make the simplification

$$\begin{aligned}\phi(y_1x_1,\ldots,y_nx_n) &= \inf_{w:\|w\|_\Sigma\le\tau}\left\langle w,\sum_{t=1}^{n}-x_ty_t\right\rangle + \mathcal{B}(y_1x_1,\ldots,y_nx_n)\\ &= -\tau\cdot\left\|\sum_{t=1}^{n}x_ty_t\right\|_\sigma + \mathcal{B}(y_1x_1,\ldots,y_nx_n).\end{aligned}$$

Hence, via the equivalence, any prediction guarantee for this setting implies a martingale inequality of the form

$$\tau\cdot\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\boldsymbol{x}_t\right\|_\sigma \le \mathop{\mathbb{E}}_{\epsilon}\mathcal{B}(\epsilon_1\boldsymbol{x}_1,\ldots,\epsilon_n\boldsymbol{x}_n),$$

for all $n\ge 1$ and all matrix-valued predictable processes $\boldsymbol{x}$.

If all $\boldsymbol{x}_t$ are indicators for the same entry in the matrix, the left-hand side of this expression can be as large as $\tau\cdot\sqrt{n}$—or $\sqrt{rd_1d_2n}$ for the prescribed choice of τ—which matches the Mirror Descent bound and gives vacuous guarantees in the worst case, thereby ruling out any uniform (constant) function $\mathcal{B}$ if we want to give a useful learning guarantee.

However, from the matrix Khintchine inequalities (Lust-Piquard and Pisier, 1991; Tropp, 2012; Mackey et al., 2014), we know that the following inequality holds for any sequence $x_1,\ldots,x_n$:

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_tx_t\right\|_\sigma \lesssim \sqrt{\left\|\sum_{t=1}^{n}x_tx_t^\top\right\|_\sigma \vee \left\|\sum_{t=1}^{n}x_t^\top x_t\right\|_\sigma}.$$

Note that this inequality only holds for individual sequences and not general martingales, but we will use it as inspiration and search for a Burkholder function for the generalized martingale inequality induced by

$$V(y_1 x_1, \ldots, y_n x_n) = \left\| \sum_{t=1}^{n} y_t x_t \right\|_\sigma - c \cdot \sqrt{\left\| \sum_{t=1}^{n} x_t x_t^\top \right\|_\sigma \vee \left\| \sum_{t=1}^{n} x_t^\top x_t \right\|_\sigma},$$

where we have set $\tau = 1$ without loss of generality, and where $c > 0$ is some constant whose value will be decided later.

We can write this expression more succinctly by introducing the Hermitian dilation and the squared Hermitian dilation (Tropp, 2012). For any matrix $X \in \mathbb{R}^{d_1 \times d_2}$ we define its Hermitian dilation $\mathcal{H}(X) \in \mathbb{S}^{d_1+d_2}$ and square $\mathcal{M}(X) \in \mathbb{S}^{d_1+d_2}$ via:

$$\mathcal{H}(X) = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix} \qquad \mathcal{M}(X) = \mathcal{H}(X)^2 = \begin{pmatrix} XX^\top & 0 \\ 0 & X^\top X \end{pmatrix}. \tag{4.9}$$

With this notation we can write

$$V(y_1 x_1, \ldots, y_n x_n) = \left\| \sum_{t=1}^{n} y_t \mathcal{H}(x_t) \right\|_\sigma - c \cdot \left\| \sum_{t=1}^{n} \mathcal{M}(x_t) \right\|_\sigma^{\frac{1}{2}}.$$

In fact, the Hermitian dilation has a symmetric spectrum, in the sense that λ is an eigenvalue of $\mathcal{H}(X)$ if and only if $-\lambda$ is an eigenvalue. This allows us to simplify to

$$V(y_1 x_1, \ldots, y_n x_n) = \lambda_{\max}\left(\sum_{t=1}^{n} y_t \mathcal{H}(x_t) \right) - c \cdot \lambda_{\max}\left(\sum_{t=1}^{n} \mathcal{M}(x_t) \right)^{\frac{1}{2}}.$$

To make finding a Burkholder function easier, we move to a relaxed version of this V function, where we introduce a new parameter $\eta > 0$ and constant $c' > 0$ and define

$$V(y_1 x_1, \ldots, y_n x_n) = \lambda_{\max}\left(\sum_{t=1}^{n} y_t \mathcal{H}(x_t) \right) - \frac{c\eta}{2} \cdot \lambda_{\max}\left(\sum_{t=1}^{n} \mathcal{M}(x_t) \right) - \frac{c'}{2\eta}.$$

We will be able to recover the original Khintchine-type guarantee using post-hoc tuning the parameter η. As a first step-toward finding a Burkholder function, we use sub-additivity of the maximum eigenvalue to write

$$V(y_1x_1,\dots,y_nx_n) \le \lambda_{\max}\left(\sum_{t=1}^n y_t\mathcal{H}(x_t) - \frac{c\eta}{2}\sum_{t=1}^n \mathcal{M}(x_t)\right) - \frac{c'}{2\eta}.$$

Now, using a standard trick in matrix concentration (Tropp, 2012), we move to the "matrix softmax" or "log-trace-exponential" function:

$$V(y_1x_1,\dots,y_nx_n) \le \frac{1}{\eta}\log \operatorname{tr}\exp\left(\eta\sum_{t=1}^n y_t\mathcal{H}(x_t) - \frac{c\eta^2}{2}\sum_{t=1}^n \mathcal{M}(x_t)\right) - \frac{c'}{2\eta}.$$

In fact, we claim that this function V is itself a Burkholder function, i.e.

$$\mathbf{U}(x_1,\dots,x_n) = \frac{1}{\eta}\log \operatorname{tr}\exp\left(\eta\sum_{t=1}^n \mathcal{H}(x_t) - \frac{c\eta^2}{2}\sum_{t=1}^n \mathcal{M}(x_t)\right) - \frac{c'}{2\eta}.$$

As a starting point, the preceding argument clearly implies that Property 1^o holds. To prove Property 3^o, we invoke Lieb's Concavity Theorem (Lieb, 1973), which states that for any fixed $A \in \mathbb{S}^d$, the function $X \mapsto \operatorname{tr}\,\exp(A + \log X)$ is concave over $\mathbb{S}^d_{++}$. Letting $x_1,\dots,x_n$ be fixed, and $S = \eta\sum_{t=1}^n \mathcal{H}(x_t) - \frac{c\eta^2}{2}\sum_{t=1}^n \mathcal{M}(x_t)$. We would like to prove that for any x,

$$\mathop{\mathbb{E}}_{\epsilon}\mathbf{U}(x_1,\dots,x_n,\epsilon x) = \mathop{\mathbb{E}}_{\epsilon}\frac{1}{\eta}\log \operatorname{tr}\exp\Big(S + \eta\epsilon\mathcal{H}(x) - \frac{c\eta}{2}\mathcal{M}(x)\Big) - \frac{c'}{2\eta} \le \frac{1}{\eta}\log \operatorname{tr}\exp(S) - \frac{c'}{2\eta}.$$

This is true via the following reasoning. Applying Lieb's concavity theorem, we have

$$\mathop{\mathbb{E}}_{\epsilon}\frac{1}{\eta}\log \operatorname{tr}\exp\Big(S + \epsilon\mathcal{H}(x) - \frac{c\eta}{2}\mathcal{M}(x)\Big) \le \frac{1}{\eta}\log \operatorname{tr}\exp\Big(S + \log\mathop{\mathbb{E}}_{\epsilon}\exp(\eta\epsilon\mathcal{H}(x)) - \frac{c\eta}{2}\mathcal{M}(x)\Big).$$

The standard matrix-valued Rademacher mgf bound (Tropp, 2012) implies that $\log\mathbb{E}_\epsilon\exp(\eta\epsilon\mathcal{H}(x)) \preceq \frac{\eta}{2}\mathcal{M}(x)$, and $A \preceq B$ implies $\operatorname{tr}e^A \le \operatorname{tr}e^B$, so the inequality indeed holds as long as $c \ge 1$.

To conclude, observe that $\log\operatorname{tr}\exp(\mathbf{0}) = \log(d_1+d_2)$, and so to ensure Property 2^o it suffices to set $c' = \log(d_1+d_2)$.

The New Algorithm What have we just accomplished? By exhibiting an explicit Burkholder function, we have just found a new efficient and adaptive algorithm for matrix prediction! Indeed, plugging the new Burkholder function into Theorem 3 has the following immediate consequence.

Corollary 1 (Burkholder Algorithm for Matrix Prediction)**.** For any fixed $\eta > 0$, the deterministic strategy

$$\widehat{y}_t = -\frac{\tau}{\eta} \cdot \mathop{\mathbb{E}}_{\sigma \in \{\pm 1\}} \left[\sigma \log \operatorname{tr} \exp \left(\eta \sigma \mathcal{H}(x_t) - \eta \sum_{s=1}^{t-1} y_s \mathcal{H}(x_s) - \tfrac{1}{2}\eta^2 \sum_{s=1}^{t} \mathcal{M}(x_s) \right) \right] \tag{4.10}$$

leads to a regret bound of

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t \leq \inf_{w : \|w\|_\Sigma \leq \tau} \sum_{t=1}^{n} -\langle w, x_t \rangle \cdot y_t + \frac{\eta \tau}{2} \cdot \left\| \sum_{t=1}^{n} x_t x_t^\top \right\|_\sigma \vee \left\| \sum_{t=1}^{n} x_t^\top x_t \right\|_\sigma + \frac{\tau \log(d_1 + d_2)}{\eta}.$$

From here it is a simple exercise to show that using a standard *doubling trick* to tune η (Cesa-Bianchi and Lugosi, 2006), yields the following stronger guarantee

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t \leq \inf_{w : \|w\|_\Sigma \leq \tau} \sum_{t=1}^{n} -\langle w, x_t \rangle \cdot y_t + \tau \cdot \sqrt{\left\| \sum_{t=1}^{n} x_t x_t^\top \right\|_\sigma \vee \left\| \sum_{t=1}^{n} x_t^\top x_t \right\|_\sigma \cdot 2 \log(d_1 + d_2)}.$$

The computation in (4.10) reduces to singular value decomposition (time complexity $O(d_1 d_2^2)$ when $d_1 \geq d_2$), and in particular does not scale with the horizon n since the method only keeps cumulative statistics in memory.

Interpreting the New Guarantee Recall that for the collaborative filtering problem, observations are incidence matrices of the form $x_t = e_{i_t} e_{j_t}^\top$. Let $N_{\text{row}} = \max_i |\{t \mid i_t = i\}|$ and $N_{\text{col}} = \max_j |\{t \mid j_t = j\}|$; these are the maximum number of times an entry appears in a given row or column, respectively. Then the bound above is equivalent to

$$\sum_{t=1}^{n} -\widehat{y}_t \cdot y_t \leq \inf_{w : \|w\|_\Sigma \leq \tau} \sum_{t=1}^{n} -\langle w, x_t \rangle \cdot y_t + \tau \cdot \sqrt{(N_{\text{col}} \vee N_{\text{row}}) \cdot 2 \log(d_1 + d_2)}.$$

As discussed earlier, to compete with the set of all rank-r matrices with bounded entries we can take $\tau = \sqrt{rd_1d_2}$. Defining $\mathrm{Reg}_n = \sum_{t=1}^n -\widehat{y}_t \cdot y_t - \inf_{w:\|w\|_\Sigma \leq \tau} \sum_{t=1}^n -\langle w, x_t\rangle \cdot y_t$, the bound above has the following favorable properties for this parameter choice:

- When entries are drawn from the uniform distribution, $N_{\text{row}} \approx n/d_1$ and $N_{\text{col}} \approx n/d_2$, which yields
$$\frac{\mathrm{Reg}_n}{n} \approx \sqrt{\frac{r(d_1 \vee d_2)}{n}}.$$
This implies that the algorithm will begin to generalize after seeing a constant number of rows worth of entries, and matches the (optimal) bound derived by Foygel and Srebro (2011) for the batch statistical learning setting.

- In general, *any* entry pattern satisfying $N_{\text{row}} \approx n/d_1$ and $N_{\text{col}} \approx n/d_2$, is sufficient to obtain the optimistic $\mathrm{Reg}_n/n \approx \sqrt{\frac{r(d_1 \vee d_2)}{n}}$ rate. Remarkably, this can happen even when the entries are chosen adaptively, so long as the condition on N_{col} and N_{row} is satisfied once the game ends.

- In the worst case, $\mathrm{Reg}_n/n \approx \sqrt{rd_1d_2/n}$, which is the minimax rate for the nuclear norm, and is obtained when the entry distribution is concentrated on a constant-sized subset of entries.

Beyond giving a new prediction algorithm, the martingale inequality implied by this Burkholder function is interesting in its own right.

Corollary 2 (Martingale Matrix Square Function Inequality)**.** For all predictable processes $\boldsymbol{x}$ and all $n \geq 1$ it holds that
$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^n \epsilon_t \boldsymbol{x}_t\right\|_\sigma \leq \sqrt{2\mathop{\mathbb{E}}_{\epsilon}\max\left\{\left\|\sum_{t=1}^n \boldsymbol{x}_t\boldsymbol{x}_t^\top\right\|_\sigma, \left\|\sum_{t=1}^n \boldsymbol{x}_t^\top\boldsymbol{x}_t\right\|_\sigma\right\}\log(d_1+d_2)}. \tag{4.11}$$

In the special case where $\boldsymbol{x}_t(\epsilon) = x_t$ is a fixed sequence, this square function inequality (4.11) recovers the Matrix Khintchine inequality (Mackey et al., 2014), including constants.

A similar martingale inequality can be obtained from the Matrix Freedman/Bennett inequalities of Tropp (2011), but this will depend on almost sure bounds on spectral norms of $(\boldsymbol{x}_t(\epsilon))_{t\leq n}$.

4.7 Discussion

The equivalence we have presented is promising both from perspective of designing efficient algorithms and from the perspective of developing fundamental limits. In the remainder of Part II we strengthen the method as follows

- In Chapter 5, we show how to deduce additional structure of the Burkholder functions from prediction inequalities of the form ϕ. This leads to a notion of sufficient statistics for online learning, and aids in the development of memory-efficient algorithms.
- In Chapter 6, we develop additional probabilistic tools based on maximal inequalities to directly prove that a given martingale inequality of the form $\mathbb{E}_{\epsilon} V \leq 0$ holds, particularly for functions V that arise in learning-theoretic applications.

We made the choice to use the linear loss $\ell(\widehat{y}, y) = -\widehat{y} \cdot y$ to simplify presentation. This shortcoming is addressed in the next chapter, and enables the use of more standard learning losses such as the absolute loss $\ell(\widehat{y}, y) = |\widehat{y} - y|$ and square loss $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$. In the general case, the form of the Burkholder algorithm is not quite as simple as in Theorem 3, but it is always efficiently computable (in terms of evaluations of $\mathbf{U}$) under mild conditions.

4.8 Chapter Notes

This chapter presents a simplified version of the results in Foster et al. (2018c).

The first work to connect the Burkholder method with online learning is (Foster et al., 2017b), which focuses on a particular application of Burkholder method related to the UMD (unconditional martingale difference) property for Banach spaces. This is covered in Chapter 8. The UMD property was the focus of the first work by Burkholder in which the method was developed (Burkholder, 1981). The present chapter is based on Foster et al. (2018c), which showed that the approach can be generalized significantly and used to address the issue of sufficient statistics for online learning (the focus of Chapter 5).

Initially introduced to give a geometric characterization of the UMD property, the Burkholder method was developed into its modern form by Burkholder in a series of works throughout the 1980s (Burkholder, 1981, 1984, 1986, 1991). Since this initial development, the design of $\mathbf{U}$ functions and related objects called Bellman functions has witnessed significant research activity in areas from harmonic analysis to optimal stopping and stochastic optimal control (Osekowski, 2012; Nazarov and Treil, 1996; Nazarov et al., 2001). The applicability to our setting has been limited so far by a focus on bounds that have sharp constants and are dimension- and horizon-independent. We anticipate that designing new $\mathbf{U}$ functions using perspectives from modern computer science, statistics, and optimization—for example, exploiting that we are tolerant to logarithmic factors in most settings—will allow us to unlock the full power of these techniques for learning applications. The "generalized martingale inequality" formulation we adopt is used throughout various works of Adam Osekowski; see Osekowski (2012) for a survey.

To the best of our knowledge, the variance-based algorithm we present for for matrix prediction is the first efficient algorithm of its kind (Arora et al., 2012; Hazan et al., 2012;

Shamir and Shalev-Shwartz, 2014; Allen-Zhu and Li, 2017). Achievability of this bound was first shown in (Foster et al., 2017b) by appealing to the UMD property for the spectral norm. The explicit construction of a UMD-style Burkholder function for the matrix prediction problem was noted to be challenging in (Foster et al., 2017b) and indeed does not appear to be known in the analysis community (Osekowski, 2017). In spite of this, the approach in this chapter uses the Burkholder method to attain the same results with an explicit (and efficient) Burkholder function.

CHAPTER 5

GENERALIZED BURKHOLDER METHOD AND SUFFICIENT STATISTICS

This chapter extends the equivalence developed in the previous chapter along two important and practical directions:

- First, we highlight substantial additional algorithmic structure exposed by the Burkholder method. We show that if an adaptive risk bound can be (approximately) expressed as a function of certain "sufficient statistics" for the data sequence, then there exists a Burkholder function that only depends on these sufficient statistics, not the entire data sequence. Following the approach of Chapter 4, this function can be used algorithmically to achieve the prediction guarantee, but it is only required to keep the sufficient statistics in memory.
- Second, we extend the techniques to handle the general online supervised learning setting (Protocol 2). The Burkholder algorithm is extended to efficiently incorporate non-linear and potentially non-smooth losses, as well as attain fast rates for nicer (e.g., strongly convex) losses. In the general case this requires solving a minimax problem at each step, which is more complicated than the closed form algorithm presented in Chapter 4, but can be carried out efficiently under mild assumptions.

We show how a many existing adaptive algorithms and prediction inequalities can be cast in the generalized Burkholder framework, and derive new efficient algorithms including the first linear-time/space prediction strategy for parameter-free supervised learning (an instance of "adaptivity to model class structure" in the language of the introduction) with linear classes and general smooth norms.

5.1 Background

Two of the most appealing features of online learning methods are (a) robustness, due to the absence of assumptions on the data-generating process, and (b) the ability to efficiently incorporate data on the fly. According to this latter desideratum, online methods should not store all the data observed so far in memory, but instead maintain some "compressed" representation, sufficient for making online predictions. The focus of this chapter is the study of such *sufficient statistics* for online learning, and the design of computationally efficient methods that employ them.

It is natural to turn to statistics for inspiration: a classical notion of *sufficient statistics* (Fisher, 1922) ensures that a statistician can search for methods that work on "compressed" representations of the data. Sufficient statistics have also been studied in sequential decision theory (Bahadur, 1954). However, the very notion of sufficiency is inherently tied to the posited probabilistic model, and the corresponding notion for arbitrary sequences—as postulated by the above desideratum (a)—is all but obvious.

The current theory of online learning offers little guidance as to what summaries of past data should be recorded by an online algorithm. For instance, the Exponential Weights algorithm (Vovk, 1990; Littlestone and Warmuth, 1994) keeps in memory the cumulative losses of the experts, while the general potential-based forecaster (Cesa-Bianchi and Lugosi, 2006) updates the cumulative regret of the algorithm with respect to each expert. The methods from the follow-the-regularized-leader family (also known as dual averaging methods) work with the sum of gradients of convex functions, while the Online Newton Step (Hazan et al., 2007) method and the Vovk-Azoury-Warmuth forecaster (Cesa-Bianchi and Lugosi, 2006) also store the "covariance" matrix of outer products. The well-known adaptive gradient descent procedure (e.g. (Rakhlin and Sridharan, 2017)) tunes the step size for online gradient descent according to the cumulative squared norms of gradients, a statistic that appears to be necessary for achieving the adaptive bound.[1]

The question of sufficient statistics for online methods appears to be unexplored and poorly understood, and it will take significant effort to answer it. In this chapter we propose an approach that appears to be general yet, inevitably, incomplete. We propose a definition that brings many existing methods under the same umbrella, and allows us to develop new efficient strategies that have otherwise been out of reach. The key workhorse for our development is the Burkholder method, following the equivalence framework developed in Chapter 4. The crucial insight is that the sufficient statistics we start with are reflected in the Burkholder function and, hence, the Burkholder algorithm is only required to update these compressed representations of the data.

5.2 Problem Setup and Sufficient Statistics

We follow the *Online Supervised Learning* setting described in Section 2.3 where, for each round $t = 1, \ldots, n$, the forecaster observes side information $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t \in \mathcal{Y} \subset \mathbb{R}$, observes an outcome $y_t \in \mathcal{Y}$, and incurs a loss of $\ell(\widehat{y}_t, y_t)$, where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In a general form, the goal of the forecaster is to ensure that

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t)\right] \leq \boldsymbol{\phi}(x_1, y_1, \ldots, x_n, y_n) \tag{5.1}$$

for any sequence $(x_1, y_1), \ldots, (x_n, y_n)$, where the expectation is with respect to forecaster's randomization. The choice of $\boldsymbol{\phi}$ of course models the problem at hand, and examples in

[1] Another example to look out for as the reader proceeds through the thesis is the ZigZag method of Chapter 8, which keeps track of a sign-transformed sequence of the gradients to achieve the empirical Rademacher complexity as a regret bound.

this chapter focus on *regret* inequalities of the form

$$\phi(x_1, y_1, \ldots, x_n, y_n) = \min_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \ell(f(x_t), y_t) + \mathcal{B}(f, x_1, \ldots, x_n) \right\}, \tag{5.2}$$

for some class of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ and an *adaptive regret bound* $\mathcal{B} : \mathcal{F} \times \mathcal{X}^n \to \mathbb{R}$.

We assume that ϕ is uniformly bounded over $(\mathcal{X} \times \mathcal{Y})^n$. We further assume that ℓ is convex and L-Lipschitz in the first argument over $\mathcal{Y}$. We denote the derivative (or a subderivative) of $\ell(\cdot, y)$ at $\widehat{y}$ by $\partial\ell(\widehat{y}, y) \in [-L, L]$. We will abbreviate $\delta_t = \partial\ell(\widehat{y}_t, y_t)$ when it is clear from context, but keep in mind that this value depends on the two variables $\widehat{y}_t$ and y_t. We assume that for any distribution p on $\mathcal{Y}$, $\arg\min_{\widehat{y} \in \mathbb{R}} \mathbb{E}_{y \sim p}\, \ell(\widehat{y}, y) \in \mathcal{Y}$, and that $\mathcal{Y}$ is compact. We let $\Delta_{\mathcal{Y}}$ denote the space of all Borel probability measures on $\mathcal{Y}$ (more generally, Δ_A will denote the set of Borel probability measures over some set A). Since $\mathcal{Y}$ is compact, Prokhorov's theorem implies that $\Delta_{\mathcal{Y}}$ is compact in the weak topology. This enables application of the minimax theorem (Theorem 1).

Additional notation For any interval $[a, b]$, we define $\mathrm{proj}_{[a,b]}(x) = \min\{b, \max\{a, x\}\}$.

5.2.1 Sufficient Statistics

Since there is no probabilistic model for data in the online learning setting, the notion of "sufficiency" has to be tied to the particular choice of adaptive rate ϕ. It is then tempting to define a sufficient statistic as a "compressed" representation which may be used by some strategy to ensure (5.1). While natural, such a definition does not provide any additional structure to narrow the search for an algorithm.

The definition we propose is as follows:

Definition 1. *Let $\mathcal{T}$ be some vector space. A function $\mathbf{T} : \mathcal{X} \times \mathcal{Y} \times [-L, L] \to \mathcal{T}$ is an* additive sufficient statistic *for ϕ if there exists $V : \mathcal{T} \to \mathbb{R}$ such that*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_1, y_1, \ldots, x_n, y_n) \le V\left(\sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t, \partial\ell(\widehat{y}_t, y_t))\right) \tag{5.3}$$

for any sequence $x_1, \widehat{y}_1, y_1, \ldots, x_n, \widehat{y}_n, y_n$. We refer to $(\mathbf{T}, V)$ as a sufficient statistic pair.

In Section 5.8, we consider a more general non-additive definition. All examples in this chapter, however, are already covered by Definition 1, and we will drop the word "additive" for now. We will also make the mild assumption that there exists $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbf{T}(x^0, y^0, 0) = 0 \in \mathcal{T}$.

Example 4 (Prediction with expert advice)**.** *Consider ϕ as in Eq. (5.2) with $\mathcal{F}$ as the set of linear functions $f(x) = \langle f, x\rangle$ for $f \in \Delta_d$, with $\mathcal{X} = [-1, 1]^d$, and with non-adaptive (uniform) rate $\mathcal{B} := c\sqrt{n \log d}$. Then the left-hand-side of (5.3) can be upper bounded via linearization of the convex loss by*

$$\max_{j\in 1,\ldots,d} \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t) \cdot (\widehat{y}_t - \langle e_j, x_t\rangle) - c\sqrt{n \log d}.$$

It follows that $\mathbb{R}^d$-valued map $\mathbf{T}$ defined by $[\mathbf{T}(x_t, \widehat{y}_t, \delta_t)]_j = \delta_t \cdot (\widehat{y}_t - \langle e_j, x_t\rangle)$ is a sufficient statistic.

Example 5 (Adaptive Gradient Descent)**.** *Consider ϕ as in Eq. (5.2) with $\mathcal{F}$ as the set of linear functions $f(x) = \langle f, x\rangle$ for $f \in \mathsf{B}_2^d$, $\mathcal{X} = \mathbb{R}^d$, and adaptive bound $\mathcal{B}(\nabla_1, \ldots \nabla_n) := (\sum_{t=1}^{n} \|\nabla_t\|^2)^{1/2}$, where $\nabla_t := \delta_t x_t$. The left-hand-side of (5.3) is at most*

$$\max_{f\in\mathsf{B}_2^d} \sum_{t=1}^{n} \delta_t \cdot (\widehat{y}_t - \langle f, x_t\rangle) - \left(\sum_{t=1}^{n} \|\nabla_t\|^2\right)^{1/2} = \sum_{t=1}^{n} \delta_t \cdot \widehat{y}_t + \left\|\sum_{t=1}^{n} \nabla_t\right\| - \left(\sum_{t=1}^{n} \|\nabla_t\|^2\right)^{1/2}. \tag{5.4}$$

This implies that $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = \left(\delta_t\widehat{y}_t, \nabla_t, \|\nabla_t\|^2\right) \in \mathbb{R} \times \mathcal{X} \times \mathbb{R}$ is a sufficient statistic.

5.3 Burkholder Method for Sufficient Statistics

The notion of sufficient statistic introduced in the previous section will only be useful if we exhibit a prediction strategy employing this representation. To do so, we introduce extensions to the Burkholder method and corresponding algorithm developed in Chapter 4.

First, we show that existence of a prediction strategy that guarantees the regret inequality (5.1) for all sequences can be ensured by checking a martingale inequality *involving only the sufficient statistics*. The key tool in proving the lemma is the minimax theorem.

Note that in a slight abuse of notation, we will concatenate the first two arguments of any sufficient statistic $\mathbf{T}$ and write them as $z_t := (x_t, \widehat{y}_t)$ going forward.

Lemma 2. Suppose $(\mathbf{T}, V)$ is a sufficient statistic pair for ϕ. Let $\delta = (\delta_1, \ldots, \delta_n)$ be a $[-L, L]$-valued martingale difference sequence (i.e. $\mathbb{E}[\delta_t \mid \mathcal{G}_{t-1}] = 0$, where $\mathcal{G}_{t-1} = \sigma(\delta_1, \ldots, \delta_{t-1})$). Let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be a sequence of functions $\boldsymbol{z}_t : [-L, L]^{t-1} \to \mathcal{X} \times \mathcal{Y}$, each viewed as a predictable process with respect to $\mathcal{G}_{t-1}$. Then a sufficient condition for existence of a prediction strategy such that (5.1) holds for all sequences $(x_1, y_1), \ldots, (x_n, y_n)$ is that

$$\mathbb{E}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \delta_t)\right)\right] \leq 0 \tag{5.5}$$

holds for any $\boldsymbol{z}$ and any law of δ. Moreover, when $\alpha \mapsto V(\tau + \mathbf{T}(z, \alpha))$ is convex for any $z \in \mathcal{X} \times \mathcal{Y}, \tau \in \mathcal{T}$, it is enough to check (5.5) for $\delta_t = \epsilon_t \cdot 2L$, $\forall t = 1, \ldots, n$, where ϵ_ts are independent Rademacher random variables.

Lemma 2 follows the results in Section 1.5 and Chapter 4 whereby existence of a strategy (or, "learnability") is certified non-constructively by proving a martingale inequality. The next lemma provides a key insight into existence of Burkholder functions with additional "geometric" properties. In particular, the Burkholder function enjoys a stronger version of the *restricted concavity* introduced in Chapter 4, in the sense that the function acts only on the space of sufficient statistics, not the entire data sequence. It is for this reason that we describe the property as "geometric". This stronger restricted concavity plays a key role in the success stories for the Burkholder method in probability, in particular for Pisier's geometric characterization of Banach spaces with the martingale type property (Pisier, 1975) and Burkholder's geometric characterization of Banach spaces with the *unconditional martingale difference* (UMD) property (Burkholder, 1981).

Lemma 3. Let $\delta = (\delta_1, \dots, \delta_n)$ be a $[-L, L]$-valued martingale difference sequence with joint law $\boldsymbol{p}$ and let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ be a predictable process ($\boldsymbol{z}_t : [-L, L]^{t-1} \to \mathcal{X} \times \mathcal{Y}$) with respect to $\mathcal{G}_{t-1} = \sigma(\delta_1, \dots, \delta_{t-1})$. The probabilistic inequality

$$\mathbb{E}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \delta_t)\right)\right] \leq 0 \tag{5.6}$$

holds for any $n \geq 1$, $\boldsymbol{z}$, and $\boldsymbol{p}$ *if and only if* one can find a function $\mathbf{U} : \mathcal{T} \to \mathbb{R}$ that satisfies the following three properties:

1^o $\mathbf{U}(0) \leq 0$.

2^o For any $\tau \in \mathcal{T}$, $\mathbf{U}(\tau) \geq V(\tau)$.

3^o For any $\tau \in \mathcal{T}$, $z \in \mathcal{X} \times \mathcal{Y}$, and any mean-zero distribution p on $[-L, L]$,

$$\mathop{\mathbb{E}}_{\alpha \sim p}\left[\mathbf{U}(\tau + \mathbf{T}(z, \alpha))\right] \leq \mathbf{U}(\tau). \tag{restricted concavity}$$

Furthermore, if for any $\tau \in \mathcal{T}$ and $z \in \mathcal{X} \times \mathcal{Y}$ the mapping $\alpha \mapsto V(\tau + \mathbf{T}(z, \alpha))$ is convex, then the condition (5.6) is implied by $\mathbb{E}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \epsilon_t \cdot 2L)\right)\right] \leq 0$, where $(\epsilon_1, \dots, \epsilon_n)$ are Rademacher random variables. For this new condition, property 3^o is replaced by

$3'$ The mapping $\alpha \mapsto \mathbf{U}(\tau + \mathbf{T}(z, \alpha))$ is convex and:

$$\forall \tau \in \mathcal{T}, z \in \mathcal{X} \times \mathcal{Y}, \quad \mathop{\mathbb{E}}_{\epsilon} \mathbf{U}(\tau + \mathbf{T}(z, \epsilon \cdot 2L)) \leq \mathbf{U}(\tau),$$

where ϵ is a Rademacher random variable.

Definition 2. *We call any function* $\mathbf{U}$ *satsifying the properties* 1^o, 2^o, *and* $3^o/3'$ *a* Burkholder function for $(\mathbf{T}, V)$.

In plain language, the lemma says that one can prove a certain probabilistic inequality if and only if there is a deterministic function with certain properties. We remark that the Burkholder functions guaranteed by the lemma are not unique, and some may be easier to find than others. We also note that any Burkholder function $\mathbf{U}$ for $(\mathbf{T}, V)$ yields another sufficient statistic pair $(\mathbf{T}, \mathbf{U})$ guaranteeing the same adaptive regret bound. The power of Theorem 3 is to guarantee the existence of a function $\mathbf{U}$ satisfying property 3^o when the function V under consideration does not have these properties. This situation, where the choice of V is "obvious" but the discovery of $\mathbf{U}$ requires nontrivial analysis, occurs frequently when one attempts to design adaptive algorithms for a new task.

To showcase the power of this lemma, we consider a particular martingale inequality that gives rise to the geometric notions of strong convexity and smoothness. These geometric properties are extensively employed in Online Convex Optimization: to instantiate the Mirror Descent algorithm with a given norm, one needs to exhibit a function that is strongly convex with respect to a given norm of interest. For example, for the ℓ_1 norm a standard choice is the negative entropy function. The next example shows that for any norm, the optimal strongly convex function is precisely the dual of the special Burkholder function for a particular martingale inequality. This example is the focus of Pisier (1975), yet for us it is one point on the spectrum of sufficient statistics.

Example 6 (Smoothness and Strong Convexity)**.** *Assume* $L = 1$ *for brevity. Suppose* $\mathcal{X} = \mathbb{R}^d$ *(more generally, we may take* $\mathcal{X}$ *to be a Banach space), equipped with a norm* $\|\cdot\|$*. Let* $V : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ *be defined by* $(x, a) \mapsto \|x\|^2 - C \cdot a$ *for* $C > 0$*. Take* $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t x_t, \|x_t\|^2)$*. Since*

$\alpha \mapsto V(\tau + \mathbf{T}(x_t, \widehat{y}_t, \alpha))$ is convex, it is enough to consider (5.6) for independent Rademacher random variables. The martingale inequality (5.6) then reads

$$\mathbb{E}\left[\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t\right\|^2 - C\sum_{t=1}^{n} \|\boldsymbol{x}_t\|^2\right] \leq 0 \tag{5.7}$$

for any $\mathcal{X}$-valued predictable process $(\boldsymbol{x}_t)$ with respect to the dyadic filtration $\mathcal{F}_{t-1} = \sigma(\epsilon_1, \ldots, \epsilon_{t-1})$. If (5.7) holds, Theorem 3 guarantees existence of a Burkholder function $\mathbf{U}$, and property $3'$ reads

$$\mathop{\mathbb{E}}_{\epsilon} \mathbf{U}(\tau_1 + \epsilon x, \tau_2 + \|x\|^2) \leq \mathbf{U}(\tau_1, \tau_2),$$

for any $\tau = (\tau_1, \tau_2) \in \mathcal{X} \times \mathbb{R}$ and $x \in \mathcal{X}$. From the construction of $\mathbf{U}$ in the proof of Theorem 3, with our particular choice of V, one can deduce that $\mathbf{U}(\tau_1, \tau_2) = \mathbf{U}(\tau_1, 0) + \tau_2$. Hence,

$$\frac{1}{2}\left(\mathbf{U}(\tau_1 + x, 0) + \mathbf{U}(\tau_1 - x, 0)\right) + C\|x\|^2 = \frac{1}{2}\left(\mathbf{U}(\tau_1 + x, C\|x\|^2) + \mathbf{U}(\tau_1 - x, C\|x\|^2)\right) \leq \mathbf{U}(\tau_1, 0)$$

and, thus, $x \mapsto \mathbf{U}(x, 0)$ is smooth with respect to the norm and its dual is strongly convex with respect to $\|\cdot\|_\star$. In summary, the Burkholder method captures the geometry necessary for defining Gradient-Descent-style methods, as the dual of $\mathbf{U}(x, 0)$ provides the universal construction for a strongly convex function with respect to a given norm. See Srebro et al. (2011) for an in-depth treatment of Mirror Descent and universal construction of strongly convex regularizers.

What should an algorithm designer take away from the developments thus far? Let us provide a brief summary. One first starts with a desired regret inequality for the online learning setting, such as (5.1). The next step is to find an upper bound on the regret inequality that can be expressed in terms of additive sufficient statistics. Lemma 2 and Theorem 3 then guarantee, respectively, that there is a certain martingale inequality that must hold if the upper bound in terms of sufficient statistics is achievable, and that there must exist a Burkholder function with certain geometric properties. In the next section we close the loop by showing that whenever such a Burkholder function can be evaluated efficiently, it yields an efficient algorithm that only keeps the sufficient statistics in memory.

Before proceeding, we briefly remark that if the sufficient statistic expansion V also serves as a lower bound on the regret inequality, then there is a formal sense in which the special Burkholder function exists if and only if there exists a strategy achieving the original regret inequality of interest; this is the focus of Section 5.7. In the reverse direction, one may start with a probabilistic inequality and determine the statistics that should be used to define the online prediction goal.[2]

5.4 Generalized Burkholder Algorithm

Example 6 in the previous section already suggests that the Burkholder $\mathbf{U}$ functions capture the "geometry" needed for forming online predictions. Indeed, the method applies to settings in which more complicated sufficient statistics (beyond the norm of the sum and the sum of the squared norms) are necessary. We now define an extension of the Burkholder algorithm from Chapter 4 based on the new concept of Burkholder functions for sufficient statistics.

To define the algorithm, first let $\zeta_{t-1} = \sum_{j=1}^{t-1} \mathbf{T}(x_j, \widehat{y}_j, \delta_j)$ be the cumulative value of the sufficient statistic computed after $t-1$ rounds. Since $\mathcal{T}$ is a vector space, ζ_ts are elements of $\mathcal{T}$, and this is the only information the algorithm stores in memory.

The *generalized Burkholder algorithm* is defined by the update:

$$\textbf{Compute } \; q_t = \operatorname*{arg\,min}_{q \in \Delta_{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y} \sim q} \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial \ell(\widehat{y}, y))\Big). \qquad \textbf{Sample } \; \widehat{y}_t \sim q_t. \tag{5.8}$$

Lemma 4. For a sufficient statistic pair $(\mathbf{T}, V)$, if there exists a Burkholder function $\mathbf{U}$ satisfying Properties 1^o, 2^o, and 3^o (or $3'$) of Theorem 3, then the Burkholder algorithm (5.8)

[2] This was precisely the approach used to develop a matrix prediction method we present in Section 5.5.4.

obtains the regret bound (5.1) in expectation for all sequences $(x_1, y_1), \ldots, (x_n, y_n)$.

Proof. To check that the above strategy works, fix a value x_t and observe that by the minimax theorem,[3]

$$\inf_{q \in \Delta_{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y} \sim q \in \Delta \mathcal{Y}} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial \ell(\widehat{y}, y))\right) = \sup_{p \in \Delta_{\mathcal{Y}}} \inf_{\widehat{y} \in \mathcal{Y}} \mathop{\mathbb{E}}_{y \sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial \ell(\widehat{y}, y))\right)$$

For any fixed p, let $\widehat{y}^\star := \arg\min_{\widehat{y} \in \mathcal{Y}} \mathbb{E}_{y \sim p} \ell(\widehat{y}, y)$, which implies $\partial \ell(\widehat{y}^\star, y)$ is a mean-zero variable (see the proof of Lemma 2). Taking the worst case value for p and choosing $\widehat{y}^\star$ as the learner's strategy for each p yields an upper bound of $\sup_{p \in \Delta_{\mathcal{Y}}} \mathbb{E}_{y \sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}^\star, \partial \ell(\widehat{y}^\star, y))\right)$, which in turn is upper bounded by

$$\sup_{\widehat{y}^\star \in \mathcal{Y}} \sup_{p \in \Delta_{[-L,L]} : \mathbb{E}_{\alpha \sim p}[\alpha] = 0} \mathop{\mathbb{E}}_{\alpha \sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}^\star, \alpha)\right)$$

by observing that the distribution over $\partial \ell(\widehat{y}^\star, y)$ belongs to the set of all zero-mean distributions supported on $[-L, L]$. The third property of $\mathbf{U}$ now leads to the upper bound,

$$\sup_{\widehat{y}^\star \in \mathcal{Y}} \sup_{p \in \Delta_{[-L,L]} : \mathbb{E}_{\alpha \sim p}[\alpha] = 0} \mathop{\mathbb{E}}_{\alpha \sim p} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}^\star, \alpha)\right) \leq \mathbf{U}\left(\zeta_{t-1}\right).$$

Applying this argument from $t = n$ down to $t = 0$ yields the value $\mathbf{U}(0) \leq 0$. □

Implementation When $\mathbf{U}$ is convex in $\widehat{y}$ and the set $\mathcal{Y}$ is convex, the minimum over q is achieved at a deterministic strategy, and so the minimization problem simplifies to $\arg\min_{\widehat{y} \in \mathcal{Y}}$. All of the Burkholder functions we explore in this chapter enjoy this or similar simplified and efficient representations for the algorithm. These simplifications are detailed in Section 5.9.1. Even without convexity, the general form for the Burkholder algorithm in (5.8) can be implemented efficiently via convex programming, assuming only Lipschitz continuity of $\mathbf{U}$.

[3] The minimax theorem can be applied because $\Delta_{\mathcal{Y}}$ is compact; see Section 2.6.

Proposition 3. Suppose $\mathbf{U}$ is Lipschitz and bounded and can be evaluated in constant time. Then (5.8) can be implemented approximately so as to achieve the regret inequality (5.1) up to additive constants in time $\text{poly}(n)$.

A precise version of this claim is deferred to Section 5.9.1.

5.5 Examples

5.5.1 ZigZag Algorithm and the UMD Property

Pisier (1975) used martingale techniques to provide a characterization of super-reflexive Banach spaces as those admitting an equivalent uniformly convex norm. As already described in Example 6, the essential ingredient of this analysis is a construction of a function $\mathbf{U}$ with the desired restricted concavity property (which turns out to be equivalent to uniform smoothness) for the martingale inequality (5.7). The corresponding notion in the world of online learning is that of an adaptive gradient (or mirror) descent.

Burkholder (1981) provided a geometrical characterization of UMD spaces, and a key ingredient of the approach was to establish existence of (and sometimes to compute in closed form) the function $\mathbf{U}$ with corresponding geometric properties (ζ-convexity, which is equivalent to "zigzag concavity" (Osekowski, 2012)). To give a teaser for Chapter 8, in the online learning world the corresponding adaptive regret bound is that of empirical Rademacher averages:

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|\leq 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - C\,\mathbb{E}\left\| \sum_{t=1}^{n} \epsilon_t \delta_t x_t \right\|.$$

By linearizing the loss, it suffices to use the sufficient statistic $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t \widehat{y}_t, \delta_t x_t, \epsilon_t x_t)$

where (ϵ_t) is taken to be a sequence drawn by the algorithm. The corresponding martingale inequality is

$$\mathbb{E}\left[\left\|\sum_{t=1}^{n}\epsilon_t\boldsymbol{x}_t(\epsilon)\right\|^p - C\left\|\sum_{t=1}^{n}\epsilon_t'\boldsymbol{x}_t(\epsilon)\right\|^p\right] \leq 0, \tag{5.9}$$

where the process in the subtracted term is decoupled and $p > 1$ is arbitrary. We refer the reader to Chapter 8 for more details.

We would like to emphasize that both smoothness/strong convexity (as in Pisier's work) and the UMD property (as in Burkholder's work) are two distinct notions with distinct sets of sufficient statistics. Since the fundamental works of Pisier and Burkholder, the so-called "Burkholder method" has been employed to prove a wide range of martingale inequalities and discover the corresponding geometric properties of the special function (Osekowski, 2012; Hytönen et al., 2016). Our contribution here is to present a unifying approach for working with arbitrary sufficient statistics in *online learning*, and to show that the Burkholder approach is in fact *algorithmic*.

5.5.2 AdaGrad and Square Function Inequalities

The Burkholder method can be used to recover efficient algorithms that obtain regret bounds in the vein of diagonal AdaGrad and full-matrix AdaGrad (Duchi et al., 2011), with optimal constants.

Define a function $\mathbf{U}_{\text{square}}(x, y) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ (Osekowski, 2005, 2012) via

$$\mathbf{U}_{\text{square}}(x, y) = \begin{cases} -\sqrt{2y^2 - \|x\|_2^2}, & y \geq \|x\|_2. \\ \|x\|_2 - 2y, & y < \|x\|_2. \end{cases}$$

$\mathbf{U}_{\text{square}}$ satisfies three properties as in Theorem 3: **1.** $\mathbf{U}_{\text{square}}(x, y) \geq \|x\|_2 - 2y$, **2.** $\mathbf{U}_{\text{square}}(x, \|x\|_2) \leq 0$, and **3.** $\mathbf{U}_{\text{square}}(x + d, \sqrt{y^2 + \|d\|_2^2}) \leq \mathbf{U}_{\text{square}}(x, y) + \langle \partial_x \mathbf{U}_{\text{square}}(x, y), d \rangle$.

This function consequently leads to two algorithms in the style of AdaGrad (Duchi et al., 2011) but with optimal constants, and which we now sketch.

The first regret bound is for ℓ_2 classes, as in full-matrix AdaGrad, and has the form

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|_2 \leq 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - 2L\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2} \leq 0.$$

The associated martingale inequality is $\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_2 \leq 2\,\mathbb{E}\sqrt{\sum_{t=1}^{n}\|\boldsymbol{x}_t(\epsilon)\|_2^2}$, which was shown to be optimal in Osekowski (2005).[4] The second regret bound is for ℓ_∞ classes, as in diagonal AdaGrad, and has the form

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|_\infty \leq 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - 2L\left\|\left(\sum_{t=1}^{n} x_t^2\right)^{1/2}\right\|_1 \leq 0,$$

where x_t^2 denotes the element-wise square. This is obtained by applying the scalar version of $\mathbf{U}_{\text{square}}$ coordinate-wise. The associated martingale inequality is $\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_1 \leq 2\,\mathbb{E}\left\|\left(\sum_{t=1}^{n} \boldsymbol{x}_t(\epsilon)^2\right)^{1/2}\right\|_1$. Both regret bounds require no prior knowledge of the range of $(x_t)_{t\leq n}$, and have runtime $O(d)$ per step.

5.5.3 Strongly Convex Losses

All of the examples we have presented so far pertain to generic Lipschitz losses, and consequently have regret that grows as $\sqrt{n}$ in the worst case. We now show that the Burkholder method can also provide *logarithmic regret* regret (Cesa-Bianchi and Lugosi, 2006; Hazan et al., 2007) for strongly convex losses. We take $\mathcal{F} = \left\{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d\right\}$ and equip this space with a regularizer $\Phi(w) = \frac{1}{2}\|w\|_2^2$. We assume that the loss $\ell(\widehat{y}, y)$ is

[4] Note that the expectation is outside the square root, so this is stronger than the ubiquitous inequality $\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_2 \leq \sqrt{\mathbb{E}\sum_{t=1}^{n}\|\boldsymbol{x}_t(\epsilon)\|_2^2}$.

ρ-strongly convex and L-Lipschitz. We adopt the shorthand $z_t = (x_t, -\widehat{y}_t)$, and our goal is to obtain a data- and comparator-dependent regret bound of the form

$$\mathcal{B}_\lambda(w; z_1, \ldots, z_n) = \Phi((w, 1)) + c \log \det\left(\rho \sum_{t=1}^{n} z_t z_t^\top + \lambda I\right) - c \log \det(\lambda I).$$

for some $c > 0$. Here we recover the classical Vovk-Azoury-Warmuth-type bound for strongly convex losses (Vovk, 1998; Azoury and Warmuth, 2001). This example is important because it shows that the Burkholder method in full generality can both obtain fast rates for curved losses and obtain bounds that jointly depend on the comparator and data. The right sufficient statistic for this problem should be familiar: In addition to storing a sum of gradients, we also store the empirical covariance $\sum_{t=1}^{n} z_t z_t^\top$. We introduce one last piece of notation: For $A \succeq 0$, $\Psi_A(w) = \frac{1}{2}\langle w, Aw \rangle$.

Proposition 4. The sufficient statistic $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = \left(\delta_t z_t, z_t z_t^\top\right) \in \mathbb{R}^{d+1} \times \mathbb{S}_+^{d+1}$ and

$$V(x, A) = \Psi^\star_{\rho A + \lambda I}(x) - c \log(\det(\rho A + \lambda I) / \det(\lambda I)) \tag{5.10}$$

forms a sufficient statistic pair for the adaptive regret bound $\mathcal{B}_\lambda$.

Theorem 4. *For the sufficient statistic pair $(\mathbf{T}, V)$ in Proposition 4, $\mathbf{U} = V$ is a Burkholder function whenever $c \geq L^2/\rho$.*

Note that for this setting the natural choice for V turned out to be a Burkholder function itself. The final runtime for the algorithm is $O(d^2)$ per step.

5.5.4 Revisiting Matrix Prediction

In this section we take another look at the matrix prediction example from Chapter 4 through the lens of sufficient statistics. Our goal is to achieve a regret inequality as in

(5.2) with a class $\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid w \in \mathcal{W}\}$, where $\mathcal{W} = \left\{w \in \mathbb{R}^{d_1 \times d_2} \mid \|w\|_\Sigma \le \tau\right\}$. Here $\langle A, B \rangle = \mathrm{tr}(AB^\top)$ is the standard matrix inner product and $\|\cdot\|_\Sigma$ denotes the nuclear norm. We also let $\|\cdot\|_\sigma$ denote the spectral norm. The loss ℓ is assumed to be L-Lipschitz and regret against a matrix $w \in \mathcal{W}$ is given by $\mathrm{Reg}_n(w) := \sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \ell(\langle w, x_t \rangle, y_t)$.

Following the approach in Chapter 4, the desired regret bound takes the form

$$\mathcal{B}_\eta(x_1, \ldots, x_n) = \frac{\eta \tau L^2}{2} \left\| \sum_{t=1}^n \mathcal{M}(x_t) \right\|_\sigma + \frac{c}{\eta}, \tag{5.11}$$

for some fixed $\eta > 0$ and constant $c > 0$. The reader might already guess that $\sum_{t=1}^n \mathcal{M}(x_t)$ should be part of the sufficient statistic. This indeed the case. The sufficient statistic takes values in $\mathcal{T} = \mathbb{R} \times \mathbb{S}^{d_1+d_2} \times \mathbb{S}_+^{d_1+d_2}$ and incorporates the matrix variance terms $\mathcal{M}(x_t)$ as follows.

Proposition 5. The pair $(\mathbf{T}, V)$ defined via $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t \cdot \widehat{y}_t, \delta_t \cdot \mathcal{H}(x_t), \mathcal{M}(x_t)) \in \mathbb{R} \times \mathbb{S}^{d_1+d_2} \times \mathbb{S}_+^{d_1+d_2}$ and

$$V(a, H, M) = a + \tau \lambda_1\big(H - \tfrac{1}{2}\eta L^2 M\big) - \frac{c}{\eta}, \tag{5.12}$$

form a sufficient statistic pair for the adaptive regret bound $\mathcal{B}_\eta$.

The construction for the Burkholder function from Chapter 4 can be summarized in the language of sufficient statistics as follows.

Theorem 5. *Define* $\mathbf{U} : \mathbb{R} \times \mathbb{S}^{d_1+d_2} \times \mathbb{S}_+^{d_1+d_2} \to \mathbb{R}$ *via*

$$\mathbf{U}(a, H, M) = a + \frac{\tau}{\eta} \log \mathrm{tr} \exp\big(\eta H - \tfrac{1}{2}\eta^2 L^2 M\big) - \frac{c}{\eta}.$$

Then $\mathbf{U}$ *is a Burkholder function, for the pair* $(\mathbf{T}, V)$ *in* (5.12) *when* $c \ge \tau \log(d_1 + d_2)$.

This more general formulation allows us to extend the matrix prediction strategy to arbitrary Lipschitz losses. The algorithm granted by the Burkholder algorithm is still quite simple due to extra convexity properties of $\mathbf{U}$; see Section 5.9.1.

Corollary 3 (Matrix prediction algorithm)**.** Suppose that $\mathcal{Y} = [-B, B]$ for some $B > 0$. Then the deterministic strategy

$$\widehat{y}_t = \mathrm{proj}_{[-B,B]}\left(-\frac{\tau}{L\eta}\mathop{\mathbb{E}}_{\sigma\in\{\pm 1\}}\left[\sigma \log \operatorname{tr} \exp\left(\eta\sigma L\mathcal{H}(x_t) + \eta\sum_{s=1}^{t-1}\delta_s\mathcal{H}(x_s) - \tfrac{1}{2}\eta^2L^2\sum_{s=1}^{t}\mathcal{M}(x_s)\right)\right]\right) \tag{5.13}$$

leads to a regret bound of

$$\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \inf_{w\in\mathcal{W}}\sum_{t=1}^{n}\ell(\langle w, x_t\rangle, y_t) \leq \tfrac{1}{2}\eta L^2\tau\left\|\sum_{t=1}^{n}\mathcal{M}(x_t)\right\|_{\sigma} + \frac{\tau\log(d_1+d_2)}{\eta}.$$

The computation in (5.13) has time complexity $O(d_1d_2^2)$ when $d_1 \geq d_2$, and does not scale with the horizon n since the method only keeps the cumulative sufficient statistics in memory.

5.6 Time-Dependent Burkholder Functions

In this section we generalize the Burkholder method to allow for a sequence of *time-dependent* Burkholder functions that satisfy the properties of Theorem 3, rather than using a single Burkholder function. This extra generality is useful when we develop adaptive regret bounds that only depend on the data sequence $x_{1:n}$,$y_{1:n}$ through the length n. As an application, we give a new family of adaptive algorithms for the problem of *parameter-free online learning* (McMahan and Orabona, 2014). This type of adaptivity falls under the umbrella of model-based adaptivity introduced in Chapter 2.

The setup is as follows: We equip the subset $\mathcal{X} \subseteq \mathbb{R}^d$ with a norm $\|\cdot\|$ and assume that $\|x_t\| \leq 1$ for all t.[5] Recall that $\|\cdot\|_\star$ will denote the dual norm. Rather than constraining the benchmark class to a compact set, we set $\mathcal{W} = \mathbb{R}^d$ and set $\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid w \in \mathcal{W}\}$.

[5]The result extends verbatim to the general Banach space case; this is only to simplify presentation.

We assume smoothness of the norm: letting $\Psi(x) = \frac{1}{2}\|x\|^2$, it holds that[6] $\Psi(x+y) \leq \Psi(x) + \langle \nabla\Psi(x), y\rangle + \frac{\beta}{2}\|y\|^2$.

To ease notational burden, we will assume the loss is 1-Lipschitz in this section. We will efficiently obtain a regret bound of the form

$$\mathrm{Reg}_n(w) \leq \mathcal{B}(w) := \|w\|_\star \sqrt{2\beta n \log\left(\sqrt{\beta} n \|w\|_\star + 1\right)} + 1 \quad \forall w \in \mathbb{R}^d \tag{5.14}$$

for any such smooth norm. We begin by stating a sufficient statistic representation for the problem. This is based on a familiar potential which has appeared in previous works on parameter-free online learning (e.g. (McMahan and Orabona, 2014)) in Hilbert spaces; we extend it to any smooth norm, then use it in the Burkholder method to provide *the first linear time/linear space algorithm for parameter-free learning with general smooth norms in online supervised learning*.

Proposition 6. Suppose we are interested in an adaptive regret bound of

$$\mathcal{B}(w) = \|w\|_\star \sqrt{2an \log\left(\frac{\sqrt{an}\|w\|_\star}{\gamma} + 1\right)} + c$$

for constants $a, \gamma, c > 0$. Then $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t \cdot \widehat{y}_t, \delta_t \cdot x_t) \in \mathbb{R} \times \mathcal{X}$ and the function

$$V(b, x) = b + \gamma \exp\left(\frac{\|x\|^2}{2an}\right) - c, \tag{5.15}$$

yield a sufficient statistic pair for the regret bound $\mathcal{B}$.

Because the regret bound we provide is not horizon independent unlike previous examples, it will be convenient to allow time-indexed Burkholder functions $(\mathbf{U}_t)_{t\leq n}$. This indexing is of purely notational convenience, as time-dependent Burkholder functions fit squarely into the algorithmic framework of Lemma 4 by enlarging $\mathcal{X}$ to $\mathcal{X} \times [n]$. Nonetheless, we

[6] Cf. Section 1.7. Our analysis extends to the general case where we instead have $\frac{1}{2}\|x\|^2 \leq \Psi(x)$ for some $\Psi \neq \frac{1}{2}\|\cdot\|^2$ and the same smoothness inequality holds, which is needed for settings such as ℓ_1/ℓ_∞.

recap the analogous properties for time-dependent Burkholder functions in the proof of the following theorem.

Theorem 6. *Suppose $c = 1$, $a = \beta$, and $\gamma = 1/\sqrt{n}$ in (5.15). Then*

$$\mathbf{U}_t(b, x) := b + \frac{1}{\sqrt{n}} \exp\left(\frac{\|x\|^2}{2\beta t} + \frac{1}{2} \sum_{s=t+1}^{n} \frac{1}{s} \right) - 1,$$

is a family of time-varying Burkholder functions satisfying 1^o, 2^o, and $3'$.

This Burkholder function immediately yields both a prediction strategy achieving (5.14) and a simple probabilistic martingale inequality. We will now state them both. Because $(\mathbf{U}_t)_{t \le n}$ satisfy additional convexity properties, the strategy is especially efficient (per Section 5.9.1 and Lemma 6).

Corollary 4. Suppose that $\mathcal{Y} = [-B, B]$ for some $B > 0$. Then the deterministic prediction strategy

$$\widehat{y}_t = \mathrm{proj}_{[-B,B]}\left(-\frac{1}{\sqrt{n}} \mathop{\mathbb{E}}_{\sigma \in \{\pm 1\}} \left[\sigma \cdot \exp\left(\frac{\left\|\sum_{s=1}^{t-1} \delta_s x_s + \sigma x_t\right\|^2}{2\beta t} + \frac{1}{2} \sum_{s=t+1}^{n} \frac{1}{s} \right) \right] \right)$$

leads to a regret bound of

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) \le \|w\|_\star \sqrt{2\beta n \log\left(\sqrt{\beta} n \|w\|_\star + 1 \right)} + 1 \quad \forall w \in \mathbb{R}^d.$$

The Burkholder function family stated above and Theorem 3 certify that $\sup \mathbb{E}[V] \le 0$. One special case of this martingale inequality is the following mgf bound for vector-valued martingales under smooth norms.

Corollary 5. Let $\boldsymbol{x}_t(\epsilon) := \boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1})$ be adapted to the filtration $\mathcal{F}_{t-1} = \sigma(\epsilon_1, \ldots, \epsilon_{t-1})$ for Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$, and let $\|\boldsymbol{x}_t\| \le 1$ almost surely, where $\|\cdot\|$ is a β-smooth norm. Then it holds that

$$\mathop{\mathbb{E}}_{\epsilon} \exp\left(\frac{\left\|\sum_{t=1}^{n} \epsilon_t \boldsymbol{x}_t(\epsilon)\right\|^2}{2\beta n} \right) \le \sqrt{n}.$$

Essentially all other approaches to parameter-free online learning (McMahan and Abernethy, 2013; McMahan and Orabona, 2014; Orabona, 2014; Orabona and Pál, 2016; Cutkosky and Boahen, 2016, 2017) only provide regret bounds of the form (5.14) in the special case where $\|\cdot\|$ is a Hilbert space. The only exception is Chapter 9, which gives an algorithm for smooth norms $\|\cdot\|$, but has time $\mathrm{poly}(n)$ per step (the results of Chapter 9 are substantially more general than the smooth norm case, however.). The independent work of (Cutkosky and Orabona, 2018) simultaneously provided an algorithm with a similar regret guarantee and computational efficiency to Theorem 6. Our Burkholder-based algorithm has running time $O(d)$ per step and only $O(d)$ memory.[7] The key ingredient to achieving this improvement was to examine a known potential through the lens of the Burkholder method.

5.7 Necessary Conditions

In Chapter 4 we saw that martingale inequalities provide necessary conditions for achievability of adaptive rates. In this section we go further and state a simple—yet powerful—result that characterizes when existence of a Burkholder function for a sufficient statistic representation pair $(\mathbf{T}, V)$ is not only sufficient, but *necessary* to obtain a particular regret bound. The fascinating implication here is that an algorithm that operates on the compressed representation $(\mathbf{T}, V)$ must exist whenever learning is possible.

Proposition 7. Let $\delta = (\delta_1, \ldots, \delta_n)$ be a $[-L, L]$-valued martingale difference sequence over filtration $\mathcal{F}_{t-1} = \sigma(\delta_1, \ldots, \delta_{t-1})$ and let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be a sequence of functions $\boldsymbol{z}_t : [-L, L]^{t-1} \to \mathcal{X} \times \mathcal{Y}$, each viewed as a predictable process with respect to $\mathcal{F}_{t-1}$. Suppose for every such $(\delta, \boldsymbol{z})$ pair there exists a randomized adversary strategy (x_t, y_t) that

[7] Technically our algorithm only applies to the online supervised learning setting, whereas the algorithm of Foster et al. (2017a) applies to the OCO setting.

guarantees, for every learner strategy $(\widehat{y}_t)_{t\leq n}$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\ell(f(x_t),y_t)-\mathcal{B}(f;x_1,\ldots,x_n)\right]\geq\mathbb{E}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right]. \tag{5.16}$$

Then, if there exists a strategy $(\widehat{y}_t)_{t\leq n}$ that achieves the regret bound $\mathcal{B}(f;x_{1:n})$, this implies that

$$\sup_{\delta,\boldsymbol{z}}\mathbb{E}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right]\leq 0.\text{[8]}$$

Consequently, the regret bound $\mathcal{B}(f;x_{1:n})$ is achievable only if there exists a Burkholder function $\mathbf{U}:\mathcal{T}\to\mathbb{R}$ that satisfies properties $1^o/2^o/3^o$of Theorem 3.

When $\alpha\mapsto V(\tau+\mathbf{T}(z,\alpha))$ is convex for any $z\in\mathcal{X}\times\mathcal{Y},\tau\in\mathcal{T}$, we only require the preceeding inequalities to hold for $\delta_t=\epsilon_t\cdot L$, $\forall t=1,\ldots,n$, where ϵ_ts are independent Rademacher random variables. In this case achievability of the regret bound $\mathcal{B}(f;x_{1:n})$ only implies existence of a Burkholder function $\mathbf{U}$ satisfying property $3'$, not 3^o.

Linear Classes At first glance the conditions of Proposition 7 may seem fairly restrictive, but it is fairly straightforward to instantiate for all the examples in this chapter. Consider the following linear setting: Take $\mathcal{X}\subseteq\mathbb{R}^d$, $\mathcal{Y}$ arbitrary, and let $\mathcal{F}$ be a linear class of the form $\{x\mapsto\langle w,x\rangle\mid w\in\mathcal{W}\}$, where $\sup_{x\in\mathcal{X},w\in\mathcal{W}}\langle w,x\rangle\leq 1$ and $\mathcal{W}$ is symmetric. Pick an arbitrary vector space $\overline{\mathcal{T}}$, let $\overline{\mathbf{T}}:\mathcal{X}\to\overline{\mathcal{T}}$ be an any featurization of the input space, and let $F:\overline{\mathcal{T}}\to\mathbb{R}$ be an arbitrary function. Our goal will be to achieve a regret bound of the form

$$\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\leq\mathcal{B}(x_{1:n}):=F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(x_t)\right). \tag{5.17}$$

Let us first consider a natural choice of V for the upper bound in this setting. Linearizing and using symmetry of $\mathcal{W}$, we have

$$\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)-\mathcal{B}(x_{1:n})\leq\sum_{t=1}^{n}\widehat{y}_t\cdot\delta_t+\sup_{w\in\mathcal{W}}\left\langle w,\sum_{t=1}^{n}\delta_t x_t\right\rangle-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(x_t)\right).$$

[8]In the more general case, if (5.16) holds up to additive slack Δ, the corresponding condition is $\sup\mathbb{E}[V]\leq\Delta$.

This means that if we choose a sufficient statistic $\mathbf{T} : (x_t, \widehat{y}_t, \delta_t) \mapsto (\widehat{y}_t\delta_t, x_t\delta_t, \overline{\mathbf{T}}(x_t)) \in \mathbb{R} \times \mathbb{R}^d \times \overline{\mathcal{T}}$ and choose $V(a, x, \overline{\tau}) = a + \sup_{w\in\mathcal{W}}\langle w, x\rangle - F(\overline{\tau})$, then it holds that

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \mathcal{B}(x_{1:n}) \leq V\left(\sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t, \delta_t)\right).$$

Noting that $\alpha \mapsto V(\tau + \mathbf{T}(x, \widehat{y}, \alpha))$ is convex, Lemma 2 implies that a sufficient condition to achieve the regret bound for any convex 1-Lipschitz loss is that

$$\sup_{\boldsymbol{z}} \mathbb{E}_{\epsilon}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \epsilon_t)\right)\right] \leq 0, \tag{5.18}$$

where $\boldsymbol{z}$ is any $\mathcal{X} \times \mathcal{Y}$-valued predictable process with respect to the Rademacher sequence $\epsilon_1, \ldots, \epsilon_n$.

By specializing to the absolute loss $\ell(\widehat{y}, y) = |\widehat{y} - y|$ and choosing an adversary that plays y_t to be Rademacher random variables and x_t to be any predictable sequence, it can be shown that (5.18) is also *necessary*; this is proven formally in Section 5.10. As a corollary, we derive the following result.

Proposition 8. There exists a Burkholder function $\mathbf{U}$ for the pair $(\mathbf{T}, V)$ *if and only if* the regret bound (5.17) is achievable.

Consider the matrix prediction setting of Section 5.5.4 for the special case of $L = 1$ and $r = 1$. This setting fits into the linear class framework above by taking $\mathcal{W}$ to be the nuclear norm ball in $\mathbb{R}^{d_1\times d_2}$ and setting $\overline{\mathbf{T}}(X) = \mathcal{M}(X)$ for any matrix $X \in \mathbb{R}^{d_1\times d_2}$. For this setting Proposition 8 implies the following equivalence.

Example 7 (Matrix Prediction)**.** *The following are equivalent:*

1. *The regret bound*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{w\,:\,\|w\|_{\Sigma}\leq 1} \sum_{t=1}^{n} \ell(\langle w, x_t\rangle, y_t) \leq \frac{\eta}{2}\left\|\sum_{t=1}^{n} \mathcal{M}(x_t)\right\|_{\sigma} + \frac{c}{\eta}$$

   is achievable.

2. *The martingale inequality*

$$\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t \boldsymbol{x}_t(\epsilon)\right\|_{\sigma} \leq \frac{\eta}{2}\mathop{\mathbb{E}}_{\epsilon}\left\|\sum_{t=1}^{n}\mathcal{M}(\boldsymbol{x}_t(\epsilon))\right\|_{\sigma} + \frac{c}{\eta}$$

holds for all $\mathbb{R}^{d_1\times d_2}$-valued predictable processes $\boldsymbol{x}$.

3. *There exists a Burkholder function for the sufficient statistic pair $(\mathbf{T}, V)$ in (5.18).*

5.8 Discussion

The core techniques developed in this chapter suggest a number of fascinating future directions which we hope will lead to a deeper understanding of online prediction and adaptive learning.

Finding sufficient statistics We gave multiple examples of Burkholder function constructions and sufficient statistics. If one wishes to find sufficient statistics for an adaptive bound $\mathcal{A}$ of interest, a basic rule of thumb is to consider a single input instance (instead of all n data points) and determine—say—a polynomial expansion or expansion in another basis for the terms in $\mathrm{Reg}_n - \mathcal{A}$ involving the instance. This gives a coarse sketch of which statistics are necessary.

As an example, take the standard square loss with linear predictors as the benchmark class and suppose we are interested in a non-adaptive bound. Following the heuristic above, we need to find an expansion for terms of the form "$(\hat{y}-y)^2 - (\langle w, x\rangle - y)^2 - \textsf{constant}$". Expanding this expression out, we find that $\hat{y}^2$, $y \cdot x$ and $xx^\top$ are all required to write the expression explicitly. In fact, for this square loss example, the weighted sum of the x_ts and the sum of the outer products $\sum_t x_t x_t^\top$ turn out to be sufficient statistics as well.

For the examples in this chapter, we exclusively considered benchmark classes $\mathcal{F}$ that were linear, which appears to have made the search for sufficient statistics easier. However, even when one considers a class $\mathcal{F}$ of non-linear functions, the approach of trying to expand the desired regret inequality (which now involves nonlinear $f \in \mathcal{F}$) around a given instance x in terms of some basis may still help to obtain an adequate sufficient statistics. Furthermore, one may enlarge the class $\mathcal{F}$ to make the sufficient statistic search easier. For instance, if we want to learn the class of boolean decision trees of depth d, we can exploit that the class can be represented by polynomials of degree d by using the discrete Fourier coefficients of the input instances up to degree d as a sufficient statistic. In summary, for non-linear classes one may still search for sufficient statistics and Burkholder functions by expressing nonlinearities (approximately) via linear combinations of higher-order terms.

Toward plug-and-play online learning A natural next step is to automatize the search for sufficient statistics and Burkholder functions. Suppose that the sufficient statistic pair $(\mathbf{T}, V)$ is fixed and all that remains is to find a Burkholder function $\mathbf{U}$. If V can be written as a polynomial of degree over sufficient statistic space $\mathcal{T}$, a natural approach is to restrict the search to Burkholder functions $\mathbf{U}$ that are themselves polynomials and relax the inequalities $1^o/2^o/3^o$ to sum-of-squares constraints (Barak and Steurer, 2014). We can then jointly search for a function $\mathbf{U}$ and a degree-d sum-of-squares proof that this function satisfies the three properties in polynomial time once the degree of $\mathbf{U}$ is fixed. As a specific example, the problem of finding the zig-zag concave Burkholder function for ℓ_p norms explored in Foster et al. (2017b) has a sufficient statistic V that is a polynomial of degree p when $p \geq 2$ is an integer.

This approach is sound in that it will never incorrectly return a function $\mathbf{U}$ that does not satisfy the three properties, but may not be complete a-priori. An interesting direction is therefore to explore whether there are conditions under which this system can indeed be made complete.

Generalized/non-additive sufficient statistics The restriction in Definition 1 that sufficient statistics combine additively can be relaxed. A more general form is as follows. First, define a *representation space* $\mathcal{T}$. The function $\mathbf{T}$ now takes the form:

$$\mathbf{T} : \mathcal{X} \times \mathcal{Y} \times [-L, L] \times \mathcal{T} \to \mathcal{T}.$$

The restricted concavity condition for $\mathbf{U}$ under this definition becomes

$$\forall z, \tau : \quad \sup_{\mathbb{E}[\alpha]=0} \mathbb{E}_{\alpha} \mathbf{U}(\mathbf{T}(z, \alpha, \tau)) \leq \mathbf{U}(\tau).$$

Properties 1^o and 2^o of Theorem 3 remain the same. This generalized notion of a sufficient statistic allows us to move beyond additive updates—$\mathbf{T}$ can multiply z with elements of $\mathcal{T}$, for example—but still restricts storage to the space $\mathcal{T}$ and is fully compatible with the Burkholder method and general algorithm framework. The generalizations of the equivalence theorem (Theorem 3) and the Burkholder algorithm (Lemma 4) for this notion of sufficient statistic hold as well.

5.9 Additional Results

5.9.1 Burkholder Algorithm Implementation

Here we show how the generalized Burkholder method can generically be implemented in polynomial time under mild assumptions on the function $\mathbf{U}$.

Generic Implementation

In this section we assume that $\mathcal{Y} = [-B, B]$ for $B > 0$ for simplicity. The only assumption we make on the form of $\mathbf{U}$ is Lipschitzness and boundedness.

Assumption 1. The are constants K_t and H_t such that the mapping

$$\widehat{y} \mapsto \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y_t))\Big)$$

is K_t-Lipschitz and bounded in magnitude by H_t for any $y_t \in \mathcal{Y}$, $x_t \in \mathcal{X}$, and ζ_{t-1} of the form $\zeta_t = \sum_{s=1}^{t} \mathbf{T}(x_s, \widehat{y}_s, \partial(\widehat{y}_s, y_s))$.

Consider the following strategy:

- Fix precision $\varepsilon_1 > 0$ and set $N = \lceil 2B/\varepsilon_1 \rceil$.
- Define control points $z_i = \min\{-B + \varepsilon_1 \cdot i, B\}$ for $0 \le i \le N$.
- Let $\widehat{\mu}_t$ be a solution to the convex program

  $$\min_{\mu \in \Delta_N} \sup_{y \in \mathcal{Y}} \sum_{i=1}^{N} \mu_i \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial\ell(z_i, y))\Big) \tag{5.19}$$

  up to additive precision ε_2.
- Sample $\widehat{y}_t \sim \widehat{\mu}_t$.

Proposition 9. Given a Burkholder function $\mathbf{U}$, the strategy above guarantees

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t)\right] - \phi(x_1, y_1, \ldots, x_n, y_n) \le \varepsilon_1 \sum_{t=1}^{n} K_t + \varepsilon_2 n.$$

That is, the regret inequality (5.1) is obtained up to additive slack controlled by ε_1 and ε_2.

Before proving the theorem, let us discuss the computational prospects of implementing this strategy. First, suppose $K_t = K$ and $H_t = H$ $\forall t \le n$. To obtain the regret inequality up to constant error it suffices to take $\varepsilon_1 = 1/Kn$ and $\varepsilon_2 = 1/n$. In this case, we have $N = O(BKn)$.

Now we must approximately solve (5.19), which is a standard finite-dimensional convex non-smooth optimization problem. There are many possible solvers; we will choose Mirror Descent (e.g. (Nemirovski et al., 1983; Nesterov, 1998; Ben-Tal and Nemirovski, 2001)) for simplicity. Let $G(\mu) = \sup_{y \in \mathcal{Y}} \sum_{i=1}^{N} \mu_i \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial\ell(z_i, y))\Big)$. Our constraint set is ℓ_1-bounded, and the boundedness assumption on $\mathbf{U}$ implies that G is H-Lipschitz with respect to the ℓ_∞ norm. In this case, Mirror Descent with the entropic regularizer (a.k.a. multiplicative weights) guarantees an ε-approximate minimizer for $G(\mu)$ after $O(H \log(N)/\varepsilon^2)$ update steps, each of which requires one evaluation of the subgradient of this function.

Evaluating the subgradient of $G(\mu)$ requires computing a supremum over $y \in \mathcal{Y}$. If $\mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial\ell(z_i, y))\Big)$ is convex with respect to y, then the supremum is obtained in $\{\pm B\}$ and so can be checked in time $O(N)$. In this case, since each Mirror Descent update takes time $O(N)$, the total complexity of the algorithm is $O(BHKn^3 \log(BKn))$.

If the supremum over $y \in \mathcal{Y}$ does not have a closed form, we can compute an approximate subgradient by taking a grid over the range $[-B, B]$ with spacing ε' and computing the $\arg\max$ over this grid by brute force. If a $O(\varepsilon)$-precision solution to the convex program is required, then it suffices to set $\varepsilon' = \varepsilon/K$ and use the approximate subgradients in the Mirror Descent scheme above. The approximate subgradient computation time is $O(KN/\varepsilon)$ in this case, since we evaluate $\sum_{i=1}^{N} \mu_i \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial\ell(z_i, y))\Big)$ once per candiate y. The final time complexity is then $O(BHK^2n^4 \log(BKn))$.

Lastly, we remark that if we replace Mirror Descent with Mirror Prox for saddle points (Nemirovski, 2004), the dependence on n in running time for the two cases above can be improved to $O(n^2)$ and $O(n^3)$ respectively.

The runtime can improved further if a regret bound of order $O(\sqrt{n})$ is sufficient, as this requires less precision.

Proof of Proposition 9. To begin, observe that since $\widehat{\mu}_t$ is an approximate solution to (5.19), it holds that

$$\sup_{y\in\mathcal{Y}}\sum_{i=1}^{N}\widehat{\mu}_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y_t))\Big)\leq\inf_{\mu\in\Delta_N}\sup_{y\in\mathcal{Y}}\sum_{i=1}^{N}\mu_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y_t))\Big)+\varepsilon_2.$$

The remainder of the proof will show that the right-hand-side above can be bounded as

$$\begin{aligned}
&\inf_{\mu\in\Delta_N}\sup_{y\in\mathcal{Y}}\sum_{i=1}^{N}\mu_i\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,z_i,\partial\ell(z_i,y_t))\Big)\\
&\leq\inf_{q\in\Delta_{\mathcal{Y}}}\sup_{y\in\mathcal{Y}}\mathop{\mathbb{E}}_{\widehat{y}\sim q}\mathbf{U}\Big(\zeta_{t-1}+\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y))\Big)+K_t\varepsilon_1\\
&\leq\mathbf{U}(\zeta_{t-1})+K_t\varepsilon_1,
\end{aligned}$$

where the second inequality follows from property 3^o of $\mathbf{U}$ and was shown in the proof of Lemma 4.

The first inequality can be seen as follows. Let $q\in\Delta_{\mathcal{Y}}$ and $y\in\mathcal{Y}$ be fixed. Let $F(z):=\mathbf{U}(\zeta_{t-1},\mathbf{T}(x_t,z,\partial\ell(z,y)))$. Since q is a Borel probability measure and F is continuous and bounded, F is integrable with respect to q:

$$\mathop{\mathbb{E}}_{\widehat{y}\sim q}\mathbf{U}(\zeta_{t-1},\mathbf{T}(x_t,\widehat{y},\partial\ell(\widehat{y},y)))=\int_{[-B,B]}F(z)dq(z).$$

Define $\mathcal{I}_1=[z_0,z_1]$ and $\mathcal{I}_i=(z_{i-1},z_i]$ for $2\leq N$. Then $\{\mathcal{I}_i\}$ form a partition of $[-B,B]$ and the integral can be approximated as

$$
\begin{aligned}
\int_{[-B,B]} F(z)dq(z) &= \sum_{i=1}^{N} \int_{\mathcal{I}_i} F(z)dq(z) \\
&\geq \sum_{i=1}^{N} \int_{\mathcal{I}_i} F(z_i)dq(z) - \sum_{i=1}^{N} \int_{\mathcal{I}_i} |F(z_i) - F(z)|dq(z) \\
&= \sum_{i=1}^{N} q(\mathcal{I}_i)F(z_i) - \sum_{i=1}^{N} \int_{\mathcal{I}_i} |F(z_i) - F(z)|dq(z) \\
&\geq \sum_{i=1}^{N} q(\mathcal{I}_i)F(z_i) - \sum_{i=1}^{N} \int_{\mathcal{I}_i} K_t\varepsilon_1 dq(z) \\
&= \sum_{i=1}^{N} q(\mathcal{I}_i)F(z_i) - K_t\varepsilon_1 \sum_{i=1}^{N} q(\mathcal{I}_i) \\
&= \sum_{i=1}^{N} q(\mathcal{I}_i)F(z_i) - K_t\varepsilon_1.
\end{aligned}
$$

Since this holds for any $q \in \Delta_{\mathcal{Y}}$ and $y \in \mathcal{Y}$, we have

$$
\begin{aligned}
&\inf_{q\in\Delta_{\mathcal{Y}}} \sup_{y\in\mathcal{Y}} \mathbb{E}_{\widehat{y}\sim q} \mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\Big) \\
&\geq \inf_{q\in\Delta_{\mathcal{Y}}} \sup_{y\in\mathcal{Y}} \sum_{i=1}^{n} q(\mathcal{I}_i)\mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial\ell(z_i, y))\Big) - K_t\varepsilon_1 \\
&= \inf_{\mu\in\Delta_N} \sup_{y\in\mathcal{Y}} \sum_{i=1}^{n} \mu_i\mathbf{U}\Big(\zeta_{t-1} + \mathbf{T}(x_t, z_i, \partial\ell(z_i, y))\Big) - K_t\varepsilon_1.
\end{aligned}
$$

$\square$

Faster Implementation under Specific Structure

In the remainder of this section we show how to implement the Burkholder algorithm for certain special cases that enable admit especially simple strategies.

Lemma 5. Suppose that the map

$$\widehat{y} \mapsto \mathbf{U}(\tau + \mathbf{T}((x, \widehat{y}), \partial(\widehat{y}, y)))$$

is convex for all y. Then the strategy

$$\widehat{y}_t = \arg\min_{\widehat{y}\in\mathcal{Y}} \sup_{y\in\mathcal{Y}} \mathbf{U}\left(\sum_{j=1}^{t-1} \zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\right) \tag{5.20}$$

achieves the value of the game in Lemma 4.

Proof of Lemma 5. This follows by reduction to the general case:

$$\begin{aligned}\inf_{\widehat{y}\in\mathcal{Y}} \sup_{y\in\mathcal{Y}} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\right) &= \inf_{q\in\Delta_{\mathcal{Y}}} \sup_{y\in\mathcal{Y}} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \mathop{\mathbb{E}}_{\widehat{y}\sim q}[\widehat{y}], \partial\ell(\mathop{\mathbb{E}}_{\widehat{y}\sim q}[\widehat{y}], y))\right) \\ &\leq \inf_{q\in\Delta_{\mathcal{Y}}} \sup_{y\in\mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y}\sim q} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}, \partial\ell(\widehat{y}, y))\right).\end{aligned}$$

The strategy in (5.20) is the minimax strategy for second expression above. The final expression is precisely the value of the Burkholder algorithm, which is controlled when $\mathbf{U}$ is a Burkholder function via Lemma 4. □

Lemma 6. Suppose that $\mathcal{Y} = [-B, B]$ for some $B > 0$. Further suppose that we can write

$$\mathbf{U}(\tau + \mathbf{T}((x, \widehat{y}), \delta)) = \widehat{y}\cdot\delta + F(\tau, x, \delta),$$

where $\delta \mapsto F(\tau, x, \delta)$ is convex for all τ, x. Then the prediction strategy

$$\widehat{y}_t = \mathrm{proj}_{[-B,B]}\left(-\frac{1}{L}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}}[\sigma F(\zeta_{t-1}, x_t, L\sigma)]\right), \tag{5.21}$$

achieves the value of the game in Lemma 4.

Proof of Lemma 6. Let $\widetilde{y}_t$ denote the unprojected version of $\widehat{y}_t$:

$$\widetilde{y}_t = -\frac{1}{L}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}}[\sigma F(\zeta_{t-1}, x_t, L\sigma)].$$

We prove the lemma by inducting backwards. Let $t \in [n]$ be fixed. We first claim that

$$\begin{aligned}\sup_{y\in\mathcal{Y}} \mathbf{U}\left(\zeta_{t-1} + \mathbf{T}(x_t, \widehat{y}_t, \partial\ell(\widehat{y}_t, y))\right) &= \sup_{y\in\mathcal{Y}}[\widehat{y}_t\cdot\partial\ell(\widehat{y}_t, y) + F(\zeta_{t-1}, x_t, \partial\ell(\widehat{y}_t, y))] \\ &\leq \sup_{y\in\mathcal{Y}}[\widetilde{y}_t\cdot\partial\ell(\widehat{y}_t, y) + F(\zeta_{t-1}, x_t, \partial\ell(\widehat{y}_t, y))].\end{aligned}$$

This holds by the assumption that $\arg\min_{\widehat{y}\in\mathbb{R}} \ell(\widehat{y}, y)$ is obtained in $[-B, B]$ for any y. The assumption implies that for any y, $\partial\ell(\widehat{y}, y) \geq 0$ for $\widehat{y} \geq B$ and $\partial\ell(\widehat{y}, y) \leq 0$ for $\widehat{y} \leq -B$. If $\widehat{y}_t \neq \widetilde{y}_t$, then either $\widehat{y}_t = B$ and $\widetilde{y}_t > B$, so that $\partial\ell(\widehat{y}_t, y)\widehat{y}_t \leq \partial\ell(\widehat{y}_t, y)\widetilde{y}_t$, or similarly $\widehat{y}_t = -B$ and $\widetilde{y}_t < -B$, which also implies $\partial\ell(\widehat{y}_t, y)\widehat{y}_t \leq \partial\ell(\widehat{y}_t, y)\widetilde{y}_t$.

Now, by the convexity assumption of the lemma, it holds that

$$
\begin{aligned}
\sup_{y\in\mathcal{Y}}[\widetilde{y}_t \cdot \partial\ell(\widehat{y}_t, y) + F(\zeta_{t-1}, x_t, \partial\ell(\widehat{y}_t, y))] &\leq \sup_{\delta\in[-L,L]}[\widetilde{y}_t \cdot \delta + F(\zeta_{t-1}, x_t, \delta)] \\
&= \max_{\sigma\in\{\pm1\}}[\widetilde{y}_t \cdot L\sigma + F(\zeta_{t-1}, x_t, L\sigma)].
\end{aligned}
$$

The choice of $\widetilde{y}_t$ guarantees that $\widetilde{y}_t \cdot L \cdot (1) + F(\zeta_{t-1}, x_t, L \cdot (1)) = \widetilde{y}_t \cdot L \cdot (-1) + F(\zeta_{t-1}, x_t, L \cdot (-1))$; this can be seen by rearranging this equality and solving for $\widetilde{y}_t$. This means that we can take $\sigma = 1$ to obtain the maximum in the expression above. Substituting in the value of $\widetilde{y}_t$ then yields

$$
\max_{\sigma\in\{\pm1\}}[\widetilde{y}_t \cdot L\sigma + F(\zeta_{t-1}, x_t, L\sigma)] = \widetilde{y}_t \cdot L \cdot (1) + F(\zeta_{t-1}, x_t, L \cdot (1)) = \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}}[F(\zeta_{t-1}, x_t, \sigma L)].
$$

Finally, we use property $3'$ of $\mathbf{U}$ and the explicit form for $\mathbf{U}$ assumed in the lemma statement to proceed back to time $t-1$:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}}[F(\zeta_{t-1}, x_t, \sigma L)] &= \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}}[\widehat{y}_t\sigma L + F(\zeta_{t-1}, x_t, \sigma L)] \\
&= \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}} \mathbf{U}(\zeta_{t-1} + \mathbf{T}((x_t, \widehat{y}_t), \sigma L)) \\
&\leq \mathbf{U}(\zeta_{t-1}).
\end{aligned}
$$

$\square$

5.9.2 Algebra of Burkholder Functions

This section contains some additional structural results about Burkholder functions that may be useful for algorithm designers.

Proposition 10. The following statements are true:

1. Given a Burkholder function $\mathbf{U}$, if we define the $X_t = \mathbf{U}(\sum_{j=1}^{t} \mathbf{T}(z_j, \delta_j))$, then for any real-valued martingale difference sequence δ_ts and predictable z_ts, $(X_t)_{t\geq 0}$ is a supermartingale with $\mathbb{E}[X_0] \leq 0$.
2. Any convex combination of Burkholder functions is a Burkholder function.
3. The minimum of a family of Burkholder functions is a Burkholder function.
4. Suppose we have a finite set A that indexes a family of functions $V_a : \mathcal{T} \to \mathbb{R}$, each of which belongs to a sufficient statistic pair $(\mathbf{T}, V_a)$ for some regret inequality of interest, and suppose each V_a has a corresponding Burkholder function $\mathbf{U}_a$. Then the following probabilistic inequality is true:
$$\mathbb{E}\left[\max_{a\in A}\left\{V_a\left(\sum_{t=1}^{n}\mathbf{T}(z_t,\delta_t)\right) - \eta n C[a]\right\}\right] \leq \frac{1}{\eta}\log|A|,$$
where $C[a] = \sup_{\tau,z,\alpha}(\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau))^2$. Note that $C \in \mathbb{R}^A$ may be thought as a sufficient statistic, though it is fixed and does not depend on instances. Furthermore, a Burkholder function $\mathbf{U} : \mathcal{T} \times \mathbb{R}^A \to \mathbb{R}$ that certifies this inequality is:
$$\mathbf{U}(\tau,\gamma) = \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta \mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta} \tag{5.22}$$

Proof of Proposition 10. The first statement follows from property 3^o of the Burkholder function $\mathbf{U}$, which immediately implies that it is a supermartingale. The second statement is trivial. To prove the third statement it suffices to verify property 3^o, which holds due to concavity of the minimum.

We now prove the fourth statement. Given a family of Burkholder functions $\{\mathbf{U}_a\}_{a\in A}$, define a new Burkholder function $\mathbf{U} : \mathcal{T} \times \mathbb{R}^A \to \mathbb{R}$ as:

$$\mathbf{U}(\tau,\gamma) = \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta}.$$

whose sufficient statistics are the original sufficient statistic of the family of V_as along with an additional $|A|$-dimensional real vector, for which one coordinate per $a \in A$ will be used to represent $C[a] = \sup_{\tau,z,\alpha}(\mathbf{U}_a(\tau + \mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau))^2$ (note that this is a vacuous statistic as it is constant for each instance). Property 3^o for $\mathbf{U}$ holds as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\alpha}\mathbf{U}\left((\tau,\gamma) + (\mathbf{T}(z,\alpha), C)\right)\\
&\quad = \frac{1}{\eta}\mathop{\mathbb{E}}_{\alpha}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \eta^2\gamma[a] - \eta^2C[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\quad \le \frac{1}{\eta}\log\left(\sum_{a\in A}\mathop{\mathbb{E}}_{\alpha}\exp\left(\eta\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \eta^2\gamma[a] - \eta^2C[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\quad = \frac{1}{\eta}\log\left(\sum_{a\in A}\mathop{\mathbb{E}}_{\alpha}\exp\left(\eta\left(\mathbf{U}_a(\tau+\mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau)\right) + \eta\mathbf{U}_a(\tau) - \eta^2\gamma[a] - \eta^2C[a]\right)\right) - \frac{\log|A|}{\eta}.
\end{aligned}$$

Now note that by property 3^o of the Burkholder functions $\{\mathbf{U}_a\}_{a\in A}$, the random variable $X_a = (\mathbf{U}_a(\tau + \mathbf{T}(z,\alpha)) - \mathbf{U}_a(\tau))$ is such that $\mathbb{E}_\alpha[X_a] \le 0$. Further from our assumption we have that $|X_a|^2 \le C[a]$. Hence, the standard mgf bound implies $\mathbb{E}_\alpha[\exp(\eta X_a)] \le \exp(\eta^2 C[a]/2)$.

$$\begin{aligned}
&\le \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) + \frac{\eta^2}{2}C[a] - \eta^2\gamma[a] - \eta^2C[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\le \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta}.
\end{aligned}$$

For property 1^o it can be seen immediately that $\mathbf{U}(0) \le 0$. Property 2^o holds via

$$\begin{aligned}
\mathbf{U}(\tau,\gamma) &= \frac{1}{\eta}\log\left(\sum_{a\in A}\exp\left(\eta\mathbf{U}_a(\tau) - \eta^2\gamma[a]\right)\right) - \frac{\log|A|}{\eta}\\
&\ge \max_{a\in A}\{\mathbf{U}_a(\tau) - \eta\gamma[a]\} - \frac{\log|A|}{\eta} \qquad \text{(softmax upper bounds max)}\\
&\ge \max_{a\in A}\{V_a(\tau) - \eta\gamma[a]\} - \frac{\log|A|}{\eta}.
\end{aligned}$$

□

We remark that one uses non-additive sufficient statistics as discussed in Section 5.8, then one can make the bound implied by the Burkholder function $\mathbf{U}$ above more data-dependent by replacing $C[a]$ with $\sup_\delta \left(\mathbf{U}_a(\tau + \mathbf{T}(z,\delta)) - \mathbf{U}_a(\tau)\right)^2$ for each a.

5.10 Detailed Proofs

5.10.1 Proofs from Section 5.3 and Section 5.4

Proof of Lemma 2. As discussed in Chapter 2, existence of a randomized strategy for (5.1) is equivalent to the following quantity being non-positive:

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta_{\mathcal{Y}}} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_1, y_1, \ldots, x_n, y_n) \right].$$

By the minimax theorem, this is equal to

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_{\mathcal{Y}}} \inf_{\widehat{y}_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \phi(x_1, y_1, \ldots, x_n, y_n) \right].$$

Following Section 2.6, we apply the minimax theorem at each time step from $t = 1, \ldots, n$ to switch the order of the learner and the adversary; briefly, our assumptions that $\mathcal{Y}$ is a compact subset of $\mathbb{R}$ and that ℓ and ϕ are bounded are sufficient. In view of (5.3), the above quantity is upper bounded by

$$\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_{\mathcal{Y}}} \inf_{\widehat{y}_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[V\left(\sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t, \partial\ell(\widehat{y}_t, y_t)) \right) \right].$$

Now, for each time t, choose the dual strategy

$$\widehat{y}_t^* := \arg\min_{\widehat{y} \in \mathcal{Y}} \mathop{\mathbb{E}}_{y_t \sim p_t} \ell(\widehat{y}, y_t),$$

so that $0 \in \partial\, \mathbb{E}_{y_t \sim p_t} \ell(\widehat{y}_t^*, y_t)$; that this is possible is implied by the assumption on the loss ℓ stated in Section 5.2. This choice implies that $\partial \ell(\widehat{y}_t^*, y_t) = \delta_t$ is a zero mean real variable conditionally on the past, i.e. $\mathbb{E}[\delta_t \mid \mathcal{G}_t] = 0$, where $\mathcal{G}_t = \sigma(\widehat{y}_{1:t-1})$. This particular choice for the $\widehat{y}_t$ in the dual game leads to the upper bound

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_{\mathcal{Y}}} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[V\left(\sum_{t=1}^{n} \mathbf{T}(x_t, \widehat{y}_t^*, \delta_t) \right) \right],$$

which is, in turn, upper bounded by

$$\left\langle\!\!\left\langle \sup_{z_t \in \mathcal{X} \times \mathcal{Y}} \sup_{p_t \in \Delta_{[-L,L]} : \mathbb{E}[\delta_t] = 0} \mathbb{E}_{\delta_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[V\left(\sum_{t=1}^{n} \mathbf{T}(z_t, \delta_t) \right) \right].$$

The last expression can be written in the functional form as

$$\sup_{\boldsymbol{z}, \boldsymbol{p}} \mathbb{E}_{\delta \sim \boldsymbol{p}} \left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right].$$

using the notation of the lemma, with the supremum over $\boldsymbol{p}$ ranging over all joint distributions on $\delta = (\delta_1, \ldots, \delta_n)$ satisfying $\mathbb{E}[\delta_t \mid \delta_{1:t-1}] = 0$ for all $t \in [n]$. The non-positivity of the latter quantity is therefore sufficient to ensure the existence of a prediction strategy satisfying (5.1). □

Proof of Theorem 3. We first establish existence of $\mathbf{U}$ under the premise of the lemma. The construction is given by

$$\mathbf{U}(\tau) = \sup_{\boldsymbol{z}, \boldsymbol{p}} \mathbb{E}_{\delta \sim \boldsymbol{p}} \left[V\left(\tau + \sum_{t \geq 1} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right]. \tag{5.23}$$

Then under the probabilistic inequality that is the premise of the lemma, it holds that

$$\mathbf{U}(0) = \sup_{\boldsymbol{z}, \boldsymbol{p}} \mathbb{E}_{\delta \sim \boldsymbol{p}} \left[V\left(\sum_{t \geq 1} \mathbf{T}(\boldsymbol{z}_t, \delta_t) \right) \right] \leq 0.$$

Next, by our assumption, $\exists z^0$ s.t. $\mathbf{T}(z^0, 0) = 0$, we can lower bound the supremum in (5.23) by considering a particular $\boldsymbol{z}$ that is constant $\boldsymbol{z}_t := z^0$ for all t, and a distribution for δ_t that only places mass on the singleton 0. This yields a lower bound

$$\mathbf{U}(\tau) \geq V(\tau).$$

To verify the third condition, observe that for any zero-mean random variable α with distribution p supported on $[-L, L]$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\alpha}\left[\mathbf{U}(\tau+\mathbf{T}(z,\alpha))\right] &= \mathop{\mathbb{E}}_{\alpha}\left[\sup_{\boldsymbol{z},\boldsymbol{p}} \mathop{\mathbb{E}}_{\delta\sim\boldsymbol{p}}\left[V\left(\tau+\mathbf{T}(z,\alpha)+\sum_t \mathbf{T}(\mathbf{z}_t,\delta_t)\right)\right]\right] \\
&\leq \sup_{\boldsymbol{z},\boldsymbol{p}} \mathop{\mathbb{E}}_{\delta\sim\boldsymbol{p}}\left[V\left(\tau+\sum_t \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right] \\
&= \mathbf{U}(\tau).
\end{aligned}$$

For the converse, assume we have a function $\mathbf{U}$ satisfying the three properties. Fix any $\boldsymbol{z}$ and $\boldsymbol{p}$ of length n. In this case, by property 2^o, the following inequality holds deterministically:

$$V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right) \leq \mathbf{U}\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right).$$

By property 3^o, we have that for any time s,

$$\mathop{\mathbb{E}}_{\delta_n} \mathbf{U}\left(\sum_{t=1}^{s} \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right) \leq \mathbf{U}\left(\sum_{t=1}^{s-1} \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right).$$

Continuing this argument all the way to $t = 0$ and using property 1^o,

$$\sup_{\boldsymbol{z},\boldsymbol{p}} \mathop{\mathbb{E}}_{\delta\sim\boldsymbol{p}}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t,\delta_t)\right)\right] \leq \mathbf{U}(0) \leq 0.$$

$\square$

5.10.2 Proofs from Section 5.5

Sketch of proofs for claims from Section 5.5.2. For the ℓ_2 result we have

$$\begin{aligned}
&\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|_2 \le 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - 2L\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2} \\
&\le \sup_{\|w\|_2 \le 1} \left\{ \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t, -\langle w, x_t \rangle) \right\} - 2L\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2} \\
&= \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \left\| \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t) x_t \right\|_2 - 2L\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2} \\
&\le \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \mathbf{U}_{\text{square}}\left(\sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t) x_t, L\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2} \right).
\end{aligned}$$

The path from here to a Burkholder function in the sense of Theorem 3 is clear given the three properties of $\mathbf{U}_{\text{square}}$ stated in the main body.

For the ℓ_∞ result, the quantity

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \min_{\|w\|_\infty \le 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) - 2L\left\| \left(\sum_{t=1}^{n} x_t^2 \right)^{1/2} \right\|_1$$

can be upper bounded by

$$\begin{aligned}
&\sup_{\|w\|_\infty \le 1} \left\{ \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t, -\langle w, x_t \rangle) \right\} - 2L\left\| \left(\sum_{t=1}^{n} x_t^2 \right)^{1/2} \right\|_1 \\
&= \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \left\| \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t) x_t \right\|_1 - 2L\left\| \left(\sum_{t=1}^{n} x_t^2 \right)^{1/2} \right\|_1 \\
&\le \sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + \sum_{i=1}^{d} \mathbf{U}_{\text{square}}\left(\sum_{t=1}^{n} \partial\ell(\widehat{y}_t, y_t) x_t[i], L\sqrt{\sum_{t=1}^{n} (x_t[i])_2^2} \right),
\end{aligned}$$

where $x_t[i]$ refers to the ith coordinate of x_t. Once again, the three properties of $\mathbf{U}_{\text{square}}$ directly lead to a valid Burkholder function $\mathbf{U}$. $\square$

Proof of Proposition 4. Let $A_n = \rho \sum_{t=1}^{n} z_t z_t^\top + \lambda I$ and $A_0 = \lambda I$. Recall that $\Psi_A(w) = \frac{1}{2}\langle w, Aw\rangle$. We begin by rewriting the desired regret bound as

$$\mathcal{B}(w; z_1, \ldots, z_n) = \lambda\Phi((w,1)) + c\log(\det(A_n)/\det(A_0))$$

for a constant $c > 0$ to be determined. With this definition, we have

$$\begin{aligned}
&\sup_{w\in\mathbb{R}^d}\{\mathrm{Reg}_n(w) - \mathcal{B}(w; z_1, \ldots, z_n)\} \\
&= \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^{n}\ell(\widehat{y}_t, y_t) - \sum_{t=1}^{n}\ell(\langle w, x_t\rangle, y_t) - \lambda\Phi((w,1))\right\} - c\log(\det(A_n)/\det(A_0))
\end{aligned}$$

Using strong convexity of ℓ:

$$\begin{aligned}
&= \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^{n}\partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t - \langle w, x_t\rangle) - \frac{\rho}{2}(\widehat{y}_t - \langle w, x_t\rangle)^2 - \lambda\Phi((w,1))\right\} - c\log(\det(A_n)/\det(A_0)) \\
&= \sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^{n}\partial\ell(\widehat{y}_t, y_t)(-\langle (w,1), z_t\rangle) - \frac{\rho}{2}(\langle (w,1), z_t\rangle)^2 - \lambda\Phi((w,1))\right\} - c\log(\det(A_n)/\det(A_0))
\end{aligned}$$

We now move to an upper bound by allowing the final coordinate of $(w,1)$ to act as a free parameter.

$$\leq \sup_{w\in\mathbb{R}^{d+1}}\left\{\sum_{t=1}^{n}\partial\ell(\widehat{y}_t, y_t)\langle w, z_t\rangle - \frac{\rho}{2}\langle w, z_t\rangle^2 - \lambda\Phi(w)\right\} - c\log(\det(A_n)/\det(A_0))$$

We can rewrite this as

$$\begin{aligned}
&\leq \sup_{w\in\mathbb{R}^{d+1}}\left\{\left\langle w, \sum_{t=1}^{n}\partial\ell(\widehat{y}_t, y_t)z_t\right\rangle - \Psi_{\rho\Sigma_n}(w) - \lambda\Phi(w)\right\} - c\log(\det(A_n)/\det(A_0)) \\
&= \sup_{w\in\mathbb{R}^{d+1}}\left\{\left\langle w, \sum_{t=1}^{n}\partial\ell(\widehat{y}_t, y_t)z_t\right\rangle - \Psi_{A_n}(w)\right\} - c\log(\det(A_n)/\det(A_0)) \\
&= \Psi^\star_{A_n}\left(\sum_{t=1}^{n}\partial\ell(\widehat{y}_t, y_t)z_t\right) - c\log(\det(A_n)/\det(A_0)).
\end{aligned}$$

This establishes that $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = \left(\delta_t z_t, z_t z_t^\top\right) \in \mathbb{R}^{d+1} \times \mathbb{S}_+^{d+1}$ is a sufficient statistic. This is because we can write

$$V(x, A) = \Psi^\star_{\rho A + \lambda I}(x) - c\log(\det(\rho A + \lambda I)/\det(A_0)).$$

and we just proved that

$$\sup_{w\in\mathbb{R}^d}\{\mathrm{Reg}_n(w)-\mathcal{B}(x_1,\ldots,x_n)\}\leq V\left(\sum_{t=1}^{n}\mathbf{T}(x_t,\widehat{y}_t,\delta_t)\right).$$

$\square$

Proof of Theorem 4. Recall that we have defined

$$\mathbf{U}(x,A)=V(x,A)=\Psi^{\star}_{A}(x)-c\log(\det(A)/\det(A_0)).$$

We verify the properties from Theorem 3. Property 2^o is immediate, and for property 1^o we have

$$\mathbf{U}(0)=\Psi^{\star}_{0+\lambda I}(0)-c\log(\det(A_0)/\det(A_0))=0.$$

We proceed to prove property 3^o. Fix $\tau=(\tau_1,\tau_2)\in\mathcal{T}=\mathbb{R}^{d+1}\times\mathbb{S}^{d+1}_{+}$ and a mean-zero distribution p over $[-L,L]$. Then we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\alpha\sim p}\mathbf{U}(\tau+\mathbf{T}(z,\alpha))&=\mathop{\mathbb{E}}_{\alpha\sim p}\Big[\Psi^{\star}_{\rho(\tau_2+zz^{\top})+\lambda I}(\tau_1+\alpha z)-c\log(\det(\rho(\tau_2+zz^{\top})+\lambda I)/\det(A_0))\Big]\\&=\mathop{\mathbb{E}}_{\alpha\sim p}\Big[\Psi^{\star}_{\rho(\tau_2+zz^{\top})+\lambda I}(\tau_1+\alpha z)\Big]-c\log(\det(\rho(\tau_2+zz^{\top})+\lambda I)/\det(A_0)).\end{aligned}$$

Let $A=\rho(\tau_2+zz^{\top})+\lambda I$ and $B=\rho\tau_2+\lambda I$. Then since $\Psi^{\star}$ is a squared Euclidean norm and α is mean-zero:

$$\mathop{\mathbb{E}}_{\alpha\sim p}[\Psi^{\star}_{A}(\tau_1+\alpha z)]\leq\Psi^{\star}_{A}(\tau_1)+\mathop{\mathbb{E}}_{\alpha\sim p}\big[\alpha^2\big\langle z,A^{-1}z\big\rangle\big]\leq\Psi^{\star}_{A}(\tau_1)+L^2\big[\alpha^2\big\langle z,A^{-1}z\big\rangle\big].$$

Also note that since $B\preceq A$, $\Psi^{\star}_{A}(\tau_1)\leq\Psi^{\star}_{B}(\tau_1)$.

To conclude, observe that we just established

$$\mathop{\mathbb{E}}_{\alpha\sim p}\mathbf{U}(\tau+\mathbf{T}(z,\alpha))\leq\Psi^{\star}_{B}(\tau_1)+L^2\big\langle z,A^{-1}z\big\rangle-c\log(\det(A)/\det(A_0)).$$

Using a standard argument (e.g. from Cesa-Bianchi and Lugosi (2006)) and using that $A = B + \rho z z^\top$:

$$\leq \Psi^\star_B(\tau_1) + \frac{L^2}{\rho}\log(\det(A)/\det(B)) - c\log(\det(A)/\det(A_0)).$$

For $c \geq L^2/\rho$, this is bounded by

$$\begin{aligned}
&\leq \Psi^\star_B(\tau_1) - c\log(\det(B)/\det(A_0)) \\
&= \mathbf{U}(\tau).
\end{aligned}$$

□

Proof of Proposition 5. Recall that $\mathcal{B}_\eta(x_1, \ldots, x_n) = \frac{\eta\tau L^2}{2}\left\|\sum_{t=1}^n \mathcal{M}(x_t)\right\|_\sigma + \frac{c}{\eta}$. Linearizing the loss with the adaptive bound as in (5.2),

$$\begin{aligned}
&\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \inf_{w\in\mathcal{W}} \ell(\langle w, x_t\rangle, y_t) - \mathcal{B}_\eta(x_1, \ldots, x_n) \\
&\leq \sup_{w\in\mathcal{W}}\left\{\sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)(\widehat{y}_t - \langle w, x_t\rangle) - \mathcal{B}_\eta(x_1, \ldots, x_n)\right\} \\
&= \sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)\widehat{y}_t + r\left\|\sum_{t=1}^n \partial\ell(\widehat{y}_t, y_t)x_t\right\|_\sigma - \mathcal{B}_\eta(x_1, \ldots, x_n).
\end{aligned}$$

We now abbreviate $\partial\ell(\widehat{y}_t, y_t) = \delta_t$ and expand out $\mathcal{B}_\eta$, yielding

$$\sum_{t=1}^n \delta_t \cdot \widehat{y}_t + \tau\left\|\sum_{t=1}^n \delta_t x_t\right\|_\sigma - \frac{\eta\tau L^2}{2}\left\|\sum_{t=1}^n \mathcal{M}(x_t)\right\|_\sigma - \frac{c}{\eta}.$$

Using the fact that $\lambda_1(\mathcal{H}(x)) = \|x\|_\sigma$, linearity of $\mathcal{H}$, and that $\mathcal{M}(x_t)$ is positive semidefinite, we write this as

$$\sum_{t=1}^n \delta_t \cdot \widehat{y}_t + \tau\lambda_1\left(\sum_{t=1}^n \delta_t \mathcal{H}(x_t)\right) - \tau\lambda_1\left(\frac{\eta L^2}{2}\sum_{t=1}^n \mathcal{M}(x_t)\right) - \frac{c}{\eta}$$

Sub-additivity of λ_1 gives a further upper bound of

$$\sum_{t=1}^n \delta_t \cdot \widehat{y}_t + \tau\lambda_1\left(\sum_{t=1}^n \delta_t \mathcal{H}(x_t) - \frac{\eta L^2}{2}\sum_{t=1}^n \mathcal{M}(x_t)\right) - \frac{c}{\eta}$$

Then $\mathbf{T}(x_t, \widehat{y}_t, \delta_t) = (\delta_t \cdot \widehat{y}_t, \delta_t \cdot \mathcal{H}(x_t), \mathcal{M}(x_t)) \in \mathbb{R} \times \mathbb{S}^{d_1+d_2} \times \mathbb{S}_{+}^{d_1+d_2}$ is a sufficient statistic. Namely, writing

$$V(a, H, M) = a + \tau\lambda_1\left(H - \frac{\eta L^2}{2}M\right) - \frac{c}{\eta},$$

our calculation shows that

$$\sup_{w\in\mathcal{W}}\{\mathrm{Reg}_n(w) - \mathcal{B}(x_1, \ldots, x_n)\} \leq V\left(\sum_{t=1}^{n}\mathbf{T}(x_t, \widehat{y}_t, \delta_t)\right).$$

□

5.10.3 Proofs from Section 5.6

Proof of Proposition 6. We define a potential function that will eventually be used in the construction of the Burkholder function $\mathbf{U}$ we provide for V. As discussed in the main body, a variant of this potential was first introduced by McMahan and Orabona (2014) for the special case of Hilbert spaces. Let $\Psi(x) = \frac{1}{2}\|x\|^2$ (not necessarily a Hilbert space norm) and define

$$F_n(x) = \gamma \exp\left(\frac{\Psi(x)}{an}\right).$$

From (McMahan and Orabona, 2014, Lemma 14), along with the additional fact that $(f(\|\cdot\|))^{\star} = f^{\star}(\|\cdot\|_{\star})$ for general dual norm pairs, it holds that

$$F_n^{\star}(w) \leq \|w\|_{\star}\sqrt{2an\log\left(\frac{\sqrt{an}\|w\|_{\star}}{\gamma} + 1\right)}.$$

This is all we need to establish the result. We proceed as follows

$$
\begin{aligned}
&\sup_{w\in\mathbb{R}^d}\{\mathrm{Reg}_n(w)-\mathcal{B}(w)\}\\
&=\sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\ell(\langle w,x_t\rangle,y_t)-\mathcal{B}(w)\right\}\\
&\leq\sup_{w\in\mathbb{R}^d}\left\{\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)(\widehat{y}_t-\langle w,x_t\rangle)-\mathcal{B}(w)\right\}\\
&=\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)\cdot\widehat{y}_t+\sup_{w\in\mathbb{R}^d}\left\{\left\langle w,\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)x_t\right\rangle-\mathcal{B}(w)\right\}
\end{aligned}
$$

Using the inequality for the potential $F_n^\star$ stated above:

$$
\leq\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)\cdot\widehat{y}_t+F_n^\star\left(\sum_{t=1}^{n}\partial\ell(\widehat{y}_t,y_t)x_t\right)-c
$$

It follows that $\mathbf{T}(x_t,\widehat{y}_t,\delta_t)=(\delta_t\cdot\widehat{y}_t,\delta_t\cdot x_t)\in\mathbb{R}\times\mathcal{X}$ is a sufficient statistic. This is because we can write

$$
V(b,x)=b+F_n^\star(x)-c.
$$

and we have just shown that

$$
\sup_{w}\{\mathrm{Reg}_n(w)-\mathcal{B}(x_1,\ldots,x_n)\}\leq V\left(\sum_{t=1}^{n}\mathbf{T}(x_t,\widehat{y}_t,\delta_t)\right).
$$

□

Proof of Theorem 6. Since $\mathbf{U}$ depends on time, we generalize the properties of Theorem 3 to

1^o $\mathbf{U}_0(0)\leq 0$

2^o For any $\tau\in\mathcal{T}$, $\mathbf{U}_n(\tau)\geq V(\tau)$

3^o For any $\tau\in\mathcal{T}$, $z\in\mathcal{X}\times\mathcal{Y}$, and any mean-zero distribution p on $[-L,L]$, and any $t\geq 1$

$$
\mathop{\mathbb{E}}_{\alpha\sim p}[\mathbf{U}_t(\tau+\mathbf{T}(z,\alpha))]\leq\mathbf{U}_{t-1}(\tau)\tag{5.24}
$$

$3'$ For any $\tau \in \mathcal{T}$, $z \in \mathcal{X} \times \mathcal{Y}$, and any $t \geq 1$,

$$\forall \tau \in \mathcal{T}, z \in \mathcal{X} \times \mathcal{Y}, \quad \mathop{\mathbb{E}}_{\epsilon} \mathbf{U}_t(\tau + \mathbf{T}(z, \epsilon L)) \leq \mathbf{U}_{t-1}(\tau),$$

where ϵ is a Rademacher random variable.

Recall that for simplicity we assume $L = 1$ and $\mathcal{X}$ is a unit ball: $\|x\| \leq 1$. Let $\Psi(x) = \frac{1}{2}\|x\|^2$, where we have assumed that β-smoothness of Ψ:

$$\Psi(x + y) \leq \Psi(x) + \langle \nabla \Psi(x), y \rangle + \frac{\beta}{2}\|y\|^2.$$

Define a family of potentials

$$F_t(x) = \gamma \exp\left(\frac{\Psi(x)}{at} + \frac{1}{2} \sum_{s=t+1}^{n} \frac{1}{s} \right)$$

and $F_0 = \gamma \exp\left(\frac{1}{2}\sum_{t=1}^{n} \frac{1}{t}\right)$. Note that F_n here is the same as in the proof of Proposition 6.

Observe that

$$\mathbf{U}_t(b, x) = b + F_t^\star(x) - c,$$

where $F_t^\star$ is as defined as in the proof of Proposition 6. We proceed to establish the three properties of $\mathbf{U}$ from Theorem 3. Property 2^o holds since $V = \mathbf{U}_n$. We will show property $3'$ first, then conclude with property 1^o. Note that $\alpha \mapsto \mathbf{U}_t(\tau + \mathbf{T}(z, \alpha))$ is convex with respect to α, and so it indeed suffices to show property $3'$.

Fix an element $\tau = (\tau_1, \tau_2) \in \mathbb{R} \times \mathcal{X} = \mathcal{T}$ of the sufficient statistic space. At time n we have

$$\mathop{\mathbb{E}}_{\epsilon}\left[\mathbf{U}_n(\tau + \mathbf{T}(z, \epsilon))\right] = \mathop{\mathbb{E}}_{\epsilon}[\tau_1 + \epsilon \cdot \widehat{y} + F_n(\tau_2 + \epsilon x_n)] - c = \tau_1 + \mathop{\mathbb{E}}_{\epsilon}[F_n(\tau_2 + \epsilon x_n)] - c.$$

To handle F_n, begin by using smoothness of Ψ:

$$\mathop{\mathbb{E}}_{\epsilon}[F_n(\tau_2 + \epsilon x_n)] = \mathop{\mathbb{E}}_{\epsilon} \exp\left(\frac{\Psi(\tau_2 + \epsilon x)}{an} \right) \leq \mathop{\mathbb{E}}_{\epsilon} \exp\left(\frac{\Psi(\tau_2) + \epsilon \langle \nabla \Psi(\tau_2), x \rangle + \frac{\beta}{2}\|x\|^2}{an} \right)$$

Using the standard Rademacher mgf bound, $\mathbb{E}_\epsilon e^{\lambda\epsilon} \le e^{\lambda^2/2}$, we upper bound the above quantity by

$$\exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}\|x\|^2}{an} + \frac{\langle \nabla\Psi(\tau_2), x\rangle^2}{2(an)^2}\right) \le \exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}\|x\|^2}{an} + \frac{\|\nabla\Psi(\tau_2)\|_\star^2\|x\|^2}{2(an)^2}\right).$$

Using the assumption $\|x\| \le 1$, we obtain an upper bound of

$$\exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}}{an} + \frac{\|\nabla\Psi(\tau_2)\|_\star^2}{2(an)^2}\right).$$

We now use a basic fact from convex analysis, namely that any β-smooth convex function f, $\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_\star^2 \le f(x) - f(y) - \langle \nabla f(y), x - y\rangle$. This yields an upper bound

$$\exp\left(\frac{\Psi(\tau_2) + \frac{\beta}{2}}{an} + \frac{\beta\Psi(\tau_2)}{(an)^2}\right)$$

Setting $a = \beta$, this is equal to

$$\exp\left(\frac{1}{\beta}\left(\frac{1}{n} + \frac{1}{n^2}\right)\Psi(\tau_2) + \frac{1}{2n}\right).$$

As a last step, observe that $\frac{1}{n} + \frac{1}{n^2} \le \frac{1}{n-1}$. Indeed,

$$\frac{1}{n} + \frac{1}{n^2} = \frac{1}{n}\left(1 + \frac{1}{n}\right) = \frac{1}{n-1}\left(1 - \frac{1}{n}\right)\left(1 + \frac{1}{n}\right) = \frac{1}{n-1}\left(1 - \frac{1}{n^2}\right) \le \frac{1}{n-1}.$$

Therefore, we have established that

$$\mathop{\mathbb{E}}_{\epsilon}[F_n(\tau_2 + \epsilon x_n)] \le \exp\left(\frac{\Psi(\tau_2)}{\beta(n-1)} + \frac{1}{2n}\right) = F_{n-1}(\tau_2),$$

and in particular $\mathbb{E}_\epsilon \mathbf{U}_n(\tau + \mathbf{T}(z,\epsilon)) \le \mathbf{U}_{n-1}(\tau)$. In fact, by folding the terms $\frac{1}{2}\sum_{s=t+1}^{n}\frac{1}{s}$—which do not depend on data—into a multiplicative constant, this argument yields, for any t and any $\|x\| \le 1$,

$$\mathop{\mathbb{E}}_{\epsilon}[F_t(\tau + \epsilon x)] \le F_{t-1}(\tau).$$

Thus, for each $t \ge 2$ we have

$$\mathop{\mathbb{E}}_{\epsilon}\left[\mathbf{U}_t(\tau + \mathbf{T}(z,\epsilon))\right] = \mathop{\mathbb{E}}_{\epsilon}[\tau_1 + \epsilon \cdot \widehat{y} + F_n(\tau_2 + \epsilon x)] - c \le \mathbf{U}_{t-1}(\tau).$$

The argument also yields (by removing unnecessary steps):

$$\mathbb{E}_{\epsilon}[F_1(0+\epsilon x)] \leq \gamma \exp\left(\frac{1}{2}\sum_{t=1}^{n}\frac{1}{t}\right) = F_0.$$

This means that

$$\mathbf{U}_0(0) = \gamma \exp\left(\frac{1}{2}\sum_{t=1}^{n}\frac{1}{t}\right) - c \leq \gamma \exp(\log(n)/2) - c.$$

We will set $\gamma = \frac{1}{\sqrt{n}}$ and $c = 1$, which yields $\mathbf{U}_0(0) \leq 0$.

□

5.10.4 Proofs from Section 5.7

Proof of Proposition 8. Recall that the regret inequality of interest is

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - F\left(\sum_{t=1}^{n} \overline{\mathbf{T}}(x_t)\right) \leq 0.$$

As sketched in Section 5.7, Lemma 2 shows that this is implied by

$$\sup_{\boldsymbol{z}} \mathbb{E}_{\epsilon}\left[V\left(\sum_{t=1}^{n} \mathbf{T}(\boldsymbol{z}_t, \epsilon_t)\right)\right] \leq 0, \tag{5.25}$$

so the remainder of this proof will focus on the opposite direction. Suppose that $\ell(\widehat{y}, y) := |\widehat{y} - y|$ is the absolute loss. We fix a Rademacher sequence $\epsilon_1, \ldots, \epsilon_n$ and a tree $\mathbf{x}$ with $\mathbf{x}_t(\epsilon) = \mathbf{x}_t(\epsilon_1, \ldots, \epsilon_{t-1})$. As a lower bound, consider a randomized adversary that plays $y_t = \epsilon_t$ and $x_t = \mathbf{x}_t(\epsilon)$. In this case the expected value of the regret inequality is

$$\mathbb{E}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, \epsilon_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t(\epsilon)), \epsilon_t) - F\left(\sum_{t=1}^{n} \overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right].$$

Observe that for any $\epsilon \in \{\pm 1\}$ we have $\ell(\widehat{y}, \epsilon) = |1 - \widehat{y}\epsilon| \geq 1 - \widehat{y}\epsilon$. Since the range of each $f \in \mathcal{F}$ lies in $[-1, 1]$, we have $\ell(f(x), \epsilon) = 1 - f(x)\epsilon$ exactly. The expected value of the regret inequality is therefore lower bounded by

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n}(1-\widehat{y}_t\epsilon_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}(1-f(\mathbf{x}_t(\epsilon))\epsilon_t)-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}(1-f(\mathbf{x}_t(\epsilon))\epsilon_t)-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{w\in\mathcal{W}}\left\langle w,\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon)\right\rangle-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\mathbf{x}_t(\epsilon),0,\epsilon_t)\right)\right].
\end{aligned}
$$

For the final step, let $\widetilde{\mathbf{y}}$ be an arbitrary $\mathcal{Y}$-valued tree $\widetilde{\mathbf{y}}_t(\epsilon)=\widetilde{\mathbf{y}}_t(\epsilon_1,\ldots,\epsilon_{t-1})$. Using the explicit form for V, we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\epsilon}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\mathbf{x}_t(\epsilon),\widetilde{\mathbf{y}}_t(\epsilon),\epsilon_t)\right)\right]&=\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n}\epsilon_t\widetilde{\mathbf{y}}_t(\epsilon)+\sup_{w\in\mathcal{W}}\left\langle w,\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon)\right\rangle-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[0+\sup_{w\in\mathcal{W}}\left\langle w,\sum_{t=1}^{n}\epsilon_t\mathbf{x}_t(\epsilon)\right\rangle-F\left(\sum_{t=1}^{n}\overline{\mathbf{T}}(\mathbf{x}_t(\epsilon))\right)\right]\\
&=\mathop{\mathbb{E}}_{\epsilon}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\mathbf{x}_t(\epsilon),0,\epsilon_t)\right)\right].
\end{aligned}
$$

Since the argument above holds for any trees $\mathbf{x}$ and $\widetilde{\mathbf{y}}$, we conclude that the regret inequality implies that

$$
\sup_{\boldsymbol{z}}\mathop{\mathbb{E}}_{\epsilon}\left[V\left(\sum_{t=1}^{n}\mathbf{T}(\boldsymbol{z}_t,\epsilon_t)\right)\right]\leq 0.
$$

for all $\mathcal{X}\times\mathcal{Y}$-valued trees.

$\square$

5.11 Chapter Notes

This chapter is adapted from Foster et al. (2018c). We thank Adam Osekowski for suggesting the example in Section 5.5.2.

Comparison With Other Algorithmic Approaches We have already shown that the Burkholder method encompasses and extends the Mirror Descent/Follow-the-Regularized-Leader family of algorithms. An alternative algorithmic approach is the *relaxation framework*, which was introduced in Rakhlin et al. (2012) and extended to handle adaptive rates in Foster et al. (2015). Compared to the relaxation framework, the present approach can handle recursions which cannot be written in the form "$\ell(\widehat{y}_t, y_t) + \mathrm{Rel}(x_{1:t}, y_{1:t})$", e.g. when the potential function depends on past forecasts $(\widehat{y}_t)$. Furthermore, the relaxation framework offers no insight into how to deduce additional structure such as sufficient statistics from the algorithm. We remark that the potential framework for online learning described in Cesa-Bianchi and Lugosi (2006), another well-known tool, is itself subsumed by the relaxation framework and is thus subsumed by the Burkholder framework as well.

CHAPTER 6

BOUNDING THE MINIMAX VALUE: PROBABILISTIC TOOLKIT

In this chapter we introduce generic tools that can be used to directly prove new martingale inequalities that arise from the equivalence framework, thereby certifying the existence of Burkholder functions and prediction strategies.

To motivate the results, recall that many of the martingale inequalities we have seen so far correspond to standard results from probability in Banach spaces and matrix concentration. How should one proceed if, via the equivalence, they encounter a new martingale inequality that is not already known to hold? What particular properties of the function V in the martingale inequality $\sup \mathbb{E}[V] \leq 0$ (cf. (5.5)) are important? If the adaptive rate of interest involves regret against a benchmark function class $\mathcal{F}$, how does the complexity of $\mathcal{F}$ influence achievability?

These questions are particularly important when we move beyond simple linear classes. If $\mathcal{F}$ is a class of neural networks, developing computationally efficient algorithms that work for every sequence may be hopeless depending on the assumptions on the structure and weights of the networks (Livni et al., 2014). This is only a computational hurdle however, and the information-theoretic question of what rates can be achieved relative to $\mathcal{F}$ is still quite interesting.

The main contribution of this chapter is to show that extensions to so-called *sequential complexity measures* (Rakhlin et al., 2010) can be used to answer the questions above by providing generic sufficient conditions under which adaptive rates can be achieved. In particular, each adaptive rate induces a set of so-called offset complexity measures, and obtaining small upper bounds on these quantities is sufficient to demonstrate achievability.

The analysis techniques we present recover and improves a wide variety of the adaptive rates, including quantile bounds (a type of model selection bound), and small loss bounds, and their second-order variants. In addition we derive a new online PAC-Bayes (McAllester, 1999) theorem that holds for countably infinite sets.

6.1 Background

Some of the significant developments in the theoretical foundations of *online learning* have been motivated by the parallel developments in the realm of *statistical learning*. In particular, this motivation has led to martingale extensions of empirical process theory, which were shown to be the "right" notions for online learnability (Rakhlin et al., 2010, 2014). Two topics, however, have remained elusive thus far in the general supervised setting: obtaining data-dependent (e.g., small loss) bounds and establishing model selection (or, oracle-type) inequalities for online learning problems. In this chapter we exploit the equivalence of online prediction guarantees and martingale inequalities to develop new techniques for addressing both these questions.

Oracle inequalities and model selection have been topics of intense research in statistics in the last two decades (Birgé and Massart, 1998; Lugosi and Nobel, 1999; Bartlett et al., 2002). Given a sequence of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ whose union is $\mathcal{M}$, one aims to derive a procedure that selects, given an i.i.d. sample of size n, an estimator $\hat{f}$ from a model $\mathcal{M}_{\hat{m}}$ that trades off bias and variance. Roughly speaking the desired oracle bound takes the form

$$\mathrm{err}(\hat{f}) \leq \inf_{m} \left\{ \inf_{f \in \mathcal{M}_m} \mathrm{err}(f) + \mathrm{pen}_n(m) \right\},$$

where $\mathrm{pen}_n(m)$ is a penalty for the model m. Such oracle inequalities are attractive because they can be shown to hold even if the overall model $\mathcal{M}$ is too large. A central idea in the proofs of such statements (and an idea that will appear throughout the present chapter) is that $\text{pen}_n(m)$ should be "slightly larger" than the fluctuations of the empirical process for the model m. It is therefore not surprising that concentration inequalities—and particularly Talagrand's celebrated inequality for the supremum of the empirical process—have played an important role in attaining oracle bounds. In order to select a good model in a data-driven manner, one establishes non-asymptotic data-dependent bounds on the fluctuations of an empirical process indexed by elements in each model (Massart, 2007).

Lifting the ideas of oracle inequalities and data-dependent bounds from statistical to online learning is not an obvious task. For one, there is no concentration inequality available, even for the simple case of sequential Rademacher complexity. (For the reader already familiar with this complexity: a change of the value of one Rademacher variable results in a change of the remaining path, and hence an attempt to use a version of a bounded difference inequality grossly fails). Luckily, as we show in this chapter, the concentration machinery is not needed and one only requires a one-sided tail inequality. This realization is motivated by work of Mendelson (2014); Liang et al. (2015); Rakhlin and Sridharan (2014). At a high level, our approach will be to develop one-sided inequalities for the suprema of certain "offset" processes, where the offset is chosen to be "slightly larger" than the complexity of the corresponding model. We then show that these offset processes determine which data-dependent adaptive rates are achievable for online learning problems, drawing strong connections to the ideas of statistical learning.

6.1.1 Framework

We work in the *Online Supervised Learning* setting from Section 2.3 where, $\mathcal{X}$ is the set of observations, $\widehat{\mathcal{Y}}$ is the space of decisions, $\mathcal{Y}$ is the set of outcomes, and $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ is the loss function. The framework is defined by the following process: For $t = 1, \ldots, n$, Nature provides input instance $x_t \in \mathcal{X}$; Learner selects prediction distribution $q_t \in \Delta(\widehat{\mathcal{Y}})$; Nature provides label $y_t \in \mathcal{Y}$, while the learner draws prediction $\hat{y}_t \sim q_t$ and suffers loss $\ell(\hat{y}_t, y_t)$.

Compared to the previous chapters in Part II we do not restrict the output space $\widehat{\mathcal{Y}}$ to be a subset of $\mathbb{R}$. Consequently, the results in this section encompass *online linear optimization* ($\mathcal{X} = \{0\}$ is a singleton set, $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ are balls in dual Banach spaces and $\ell(\hat{y}, y) = \langle \hat{y}, y \rangle$). Recall that an *adaptive regret bound* has the form $\mathcal{B}(f; x_{1:n}, y_{1:n})$, and is said to be achievable if there exists a randomized algorithm for selecting $\hat{y}_t$ such that

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \leq \mathcal{B}(f; x_{1:n}, y_{1:n}) \quad \forall x_{1:n}, y_{1:n}, \ \forall f \in \mathcal{F}. \tag{6.1}$$

For uniform rates $\mathcal{B}$, the sequential Rademacher complexity of $\mathcal{F}$ is one of the tightest achievable uniform rates for many loss functions (Rakhlin et al., 2010; Rakhlin and Sridharan, 2014). We will show that offset versions of martingale processes that generalize the sequential Rademacher complexity provide necessary and sufficient conditions for achievability of *adaptive rates*.

We distinguish between three types of adaptive rates, according to whether $\mathcal{B}(f; x_{1:n}, y_{1:n})$ depends only on f, only on $(x_{1:n}, y_{1:n})$, or on both quantities. Whenever $\mathcal{B}$ depends on f, an adaptive regret bound can be viewed as an oracle inequality which penalizes each f according to a measure of its complexity (e.g. the complexity of the smallest model to which it belongs). As in statistical learning, an oracle inequality (6.1) may be proved for certain functions $\mathcal{B}(f; x_{1:n}, y_{1:n})$ even if a uniform bound cannot hold for any nontrivial $\mathcal{B}$.

Related Adaptive Regret Bounds The tools we introduce recover the vast majority of known adaptive rates in literature, including variance bounds, quantile bounds, localization-based bounds, and fast rates for small losses. It should be noted that while existing literature on adaptive online learning has focused on simple hypothesis classes such as finite experts and finite-dimensional p-norm balls, our results extend to general hypothesis classes, including large nonparametric ones discussed in Rakhlin and Sridharan (2014).

The case when $\mathcal{B}(f; x_{1:n}, y_{1:n}) = \mathcal{B}(x_{1:n}, y_{1:n})$ does not depend on f has received the most attention in literature. The focus is on bounds that can be tighter for "nice sequences," yet maintain near-optimal worst-case guarantees. Results of this type include Hazan and Kale (2010); Chiang et al. (2012); Duchi et al. (2011); Rakhlin and Sridharan (2013), couched in the setting of online linear/convex optimization, and Cesa-Bianchi et al. (2007) in the experts setting.

A bound of type $\mathcal{B}(f)$ was studied in Chaudhuri et al. (2009), which presented an algorithm that competes with all experts simultaneously, but with varied regret with respect to each of them depending on the quantile of the expert. Another bound of this type was given by McMahan and Orabona (2014), who consider online linear optimization with an unbounded set and provide oracle inequalities with an appropriately chosen function $\mathcal{B}(f)$.

Finally, the third category of adaptive bounds are those that depend on both the hypothesis $f \in \mathcal{F}$ and the data. The bounds that depend on the loss of the best function (so-called "small-loss" bounds, (Cesa-Bianchi and Lugosi, 2006, Sec. 2.4), Srebro et al. (2010); Cesa-Bianchi et al. (2007)) fall in this category trivially, since one may over-bound the loss of the best function by the performance of f. We draw attention to a result of Luo and Schapire (2015) who show an adaptive bound in terms of both the loss of comparator and the KL divergence between the comparator and some pre-fixed prior distribution over experts. An MDL-style bound in terms of the variance of the loss of the comparator (under the distribution induced by the algorithm) was given in Koolen and van Erven (2015).

Beyond the examples given in this chapter, the tools we develop here will be used to derive new adaptive learning guarantees in Part III.

6.2 Adaptive Rates and Achievability: General Setup

Recall from Chapter 2 that for any adaptive regret bound $\mathcal{B}$, minimax achievability is defined by

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B}) := \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \inf_{q_t\in\Delta(\widehat{\mathcal{Y}})} \sup_{y_t\in\mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\ell(f(x_t),y_t) + \mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right].$$

$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B})$ quantifies how $\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\ell(f(x_t),y_t) + \mathcal{B}(f;x_{1:n},y_{1:n})\right\}$ behaves when the optimal learning algorithm that minimizes this difference is used against Nature trying to maximize it. Directly from this definition, **An adaptive rate $\mathcal{B}$ is achievable if and only if $\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B}) \le 0$.**

If $\mathcal{B}$ is a uniform rate, i.e., $\mathcal{B}(f;x_{1:n},y_{1:n}) = \mathcal{B}$, achievability reduces to the minimax analysis explored in Rakhlin et al. (2010). The uniform rate $\mathcal{B}$ is achievable if and only if $\mathcal{B} \ge \mathcal{V}_n(\mathcal{F})$, where $\mathcal{V}_n(\mathcal{F})$ is the minimax value of the online learning game.

The aim of this chapter is to develop an understanding of when the minimax value $\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B})$ for general adaptive rates $\mathcal{B}$. We first show that the minimax value is bounded by an offset version of the sequential Rademacher complexity studied in Rakhlin et al. (2010). The symmetrization lemma (Lemma 7) below provides us with the first step towards a probabilistic analysis of achievable rates. Before stating the lemma, we need to define the notion of a tree and the notion of sequential Rademacher complexity.

Given a set $\mathcal{Z}$, a $\mathcal{Z}$-valued tree (or, predictable process) $\boldsymbol{z}$ of depth n is a sequence $(\boldsymbol{z}_t)_{t=1}^{n}$ of functions $\boldsymbol{z}_t : \{\pm 1\}^{t-1} \to \mathcal{Z}$. For a tree $\boldsymbol{z}$, the sequential Rademacher complexity of a function class $\mathcal{G} \subseteq (\mathcal{Z} \to \mathbb{R})$ on $\boldsymbol{z}$ is defined as

$$\mathcal{R}_n^{\text{seq}}(\mathcal{G}, \boldsymbol{z}) := \mathop{\mathbb{E}}_{\epsilon} \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) \quad \text{ and } \quad \mathcal{R}_n^{\text{seq}}(\mathcal{G}) := \sup_{\boldsymbol{z}} \mathcal{R}_n^{\text{seq}}(\mathcal{G}, \boldsymbol{z}) \ .$$

Lemma 7. For any lower semi-continuous loss ℓ, and any adaptive rate $\mathcal{B}$,

$$\mathcal{V}_n^{\text{ol}}(\mathcal{F}, \mathcal{B}) \leq \sup_{\mathbf{x}, \mathbf{y}, \mathbf{y}'} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ 2 \sum_{t=1}^{n} \epsilon_t \ell(f(\mathbf{x}_t(\epsilon)), \mathbf{y}_t(\epsilon)) - \mathcal{B}(f; \mathbf{x}_{1:n}(\epsilon), \mathbf{y}'_{2:n+1}(\epsilon)) \right\} \right]. \tag{6.2}$$

If one considers the supervised learning problem where $\mathcal{F} : \mathcal{X} \to \mathbb{R}$, $\mathcal{Y} \subset \mathbb{R}$ and $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss that is convex and L-Lipschitz in its first argument, then for any adaptive rate $\mathcal{B}$,

$$\mathcal{V}_n^{\text{ol}}(\mathcal{F}, \mathcal{B}) \leq \sup_{\mathbf{x}, \mathbf{y}} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ 2L \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) - \mathcal{B}(f; \mathbf{x}_{1:n}(\epsilon), \mathbf{y}_{1:n}(\epsilon)) \right\} \right]. \tag{6.3}$$

The above lemma tells us that to check whether an adaptive rate is achievable, it is sufficient to check that the corresponding adaptive sequential complexity measures are non-positive. Of course, if the above complexities are bounded by some positive quantity of a smaller order, one can form a new achievable rate $\mathcal{B}'$ by adding the positive quantity to $\mathcal{B}$.

6.3 Probabilistic Tools

The analysis in this chapter rests on certain one-sided probabilistic inequalities. We now state the first building block: a rather straightforward maximal inequality.

Proposition 11. Let $I = \{1, \ldots, N\}$, $N \leq \infty$, be a set of indices and let $(X_i)_{i \in I}$ be a sequence of random variables satisfying the following tail condition: for any $\tau > 0$,

$$\mathbb{P}(X_i - B_i > \tau) \leq C_1 \exp\left(-\tau^2/(2\sigma_i^2)\right) + C_2 \exp\left(-\tau s_i\right) \tag{6.4}$$

for some positive sequence (B_i), nonnegative sequence (σ_i) and nonnegative sequence (s_i) of numbers, and for constants $C_1, C_2 \geq 0$. Then for any $\bar{\sigma} \leq \sigma_1$, $\bar{s} \geq s_1$, and

$$\theta_i = \max\left\{ \frac{\sigma_i}{B_i} \sqrt{2 \log(\sigma_i/\bar{\sigma}) + 4 \log(i)}, (B_i s_i)^{-1} \log\left(i^2 (\bar{s}/s_i)\right) \right\} + 1,$$

it holds that

$$\mathbb{E} \sup_{i \in I} \{X_i - B_i \theta_i\} \le 3C_1 \bar{\sigma} + 2C_2 (\bar{s})^{-1}. \tag{6.5}$$

We remark that B_i need not be the expected value of X_i, as we are not interested in two-sided deviations around the mean.

A standard approach to obtain oracle-type inequalities (Massart, 2007) is to split a large class into smaller ones according to a "complexity radius" and control a certain stochastic process separately on each subset (also known as the *peeling* technique). In the applications that follow, X_i will often stand for the (random) supremum of this process on subset i, and B_i will be an upper bound on its typical size. Given deviation bounds for X_i above B_i, the dilated size $B_i \theta_i$ then allows one to pass to maximal inequalities (6.5) and thus verify achievability in Lemma 7. The same strategy works for obtaining data-dependent bounds, where we first prove tail bounds for a given size of the data-dependent quantity, then appeal to (6.5).

A simple yet powerful example for the control of the supremum of a stochastic process is an inequality due to Pinelis (Pinelis, 1994) for the norm (which is a supremum over the dual ball) of a martingale in a 2-smooth Banach space. Here we state a version of this result that can be found in Rakhlin et al. (2011), Appendix A.

Lemma 8. Let $\mathcal{Z}$ be a unit ball in a separable $(2, D)$-smooth Banach space $\mathfrak{B}$.[1] For any $\mathcal{Z}$-valued tree $\boldsymbol{z}$, and any $n > \tau / 4D^2$

$$\mathbb{P}\left(\left\| \sum_{t=1}^{n} \epsilon_t \boldsymbol{z}_t(\epsilon) \right\| \ge \tau \right) \le 2 \exp\left(-\frac{\tau^2}{8D^2 n} \right)$$

When the class of functions is not linear, we may no longer appeal to the above lemma. Instead, we make use of a result from Rakhlin et al. (2015) that extends Lemma 8 at a

[1] Cf. Section 1.7.

price of a poly-logarithmic factor. Before stating this lemma, we recall the definition of the *sequential covering number*. First, a set V of $\mathbb{R}$-valued trees is called an α-cover of $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ on $\boldsymbol{z}$ with respect to ℓ_p if

$$\forall g \in \mathcal{G}, \forall \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad \sum_{t=1}^{n} (g(\boldsymbol{z}_t(\epsilon)) - \boldsymbol{v}_t(\epsilon))^p \leq n\alpha^p.$$

The size of the smallest α-cover is denoted by $\mathcal{N}_p(\mathcal{G}, \alpha, \boldsymbol{z})$, and $\mathcal{N}_p(\mathcal{G}, \alpha, n) := \sup_{\boldsymbol{z}} \mathcal{N}_p(\mathcal{G}, \alpha, \boldsymbol{z})$.

The set V is an α-cover of $\mathcal{G}$ on $\boldsymbol{z}$ with respect to ℓ_∞ if

$$\forall g \in \mathcal{G}, \forall \epsilon \in \{\pm 1\}, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad |g(\boldsymbol{z}_t(\epsilon)) - \boldsymbol{v}_t(\epsilon)| \leq \alpha \quad \forall t \in [n].$$

We let $\mathcal{N}_\infty(\mathcal{G}, \alpha, \boldsymbol{z})$ be the smallest such cover and set $\mathcal{N}_\infty(\mathcal{G}, \alpha, n) = \sup_{\boldsymbol{z}} \mathcal{N}_\infty(\mathcal{G}, \alpha, \boldsymbol{z})$.

Lemma 9 (Rakhlin et al. (2015)). Let $\mathcal{G} \subseteq [-1, 1]^{\mathcal{Z}}$. Suppose $\mathcal{R}_n^{\text{seq}}(\mathcal{G})/n \to 0$ with $n \to \infty$ and that the following mild assumptions hold: $\mathcal{R}_n^{\text{seq}}(\mathcal{G}) \geq 1/n$, $\mathcal{N}_\infty(\mathcal{G}, 2^{-1}, n) \geq 4$, and there exists a constant Γ such that $\Gamma \geq \sum_{j=1}^{\infty} \mathcal{N}_\infty(\mathcal{G}, 2^{-j}, n)^{-1}$. Then for any $\theta > \sqrt{12/n}$, for any $\mathcal{Z}$-valued tree $\boldsymbol{z}$ of depth n,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g \in \mathcal{G}} \left| \sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) \right| > 8\left(1 + \theta\sqrt{8n \log^3(en^2)}\right) \cdot \mathcal{R}_n^{\text{seq}}(\mathcal{G})\right) \\
&\quad \leq \mathbb{P}\left(\sup_{g \in \mathcal{G}} \left| \sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) \right| > n \inf_{\alpha > 0} \left\{4\alpha + 6\theta \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_\infty(\mathcal{G}, \delta, n)} d\delta\right\}\right) \leq 2\Gamma e^{-\frac{n\theta^2}{4}}.
\end{aligned}$$

This lemma yields a one-sided control on the size of the supremum of the sequential Rademacher process, as required for our oracle-type inequalities.

Next, we turn our attention to an offset Rademacher process, where the supremum is taken over a collection of negative-mean random variables. The behavior of this offset process was shown to govern the optimal rates of convergence for online nonparametric regression (Rakhlin and Sridharan, 2014). Such one-sided control of the supremum will be necessary for some of the data-dependent upper bounds we develop.

Lemma 10. Let $\boldsymbol{z}$ be a $\mathcal{Z}$-valued tree of depth n, and let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$. For any $\gamma \geq 1/n$ and $\alpha > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\left(\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\alpha g^2(\boldsymbol{z}_t(\epsilon))\right) - \frac{\log\mathcal{N}_2(\mathcal{G},\gamma,\boldsymbol{z})}{\alpha} - 12\sqrt{2}\int_{1/n}^{\gamma}\sqrt{n\log\mathcal{N}_2(\mathcal{G},\delta,\boldsymbol{z})}d\delta - 1 > \tau\right)$$
$$\leq \Gamma\exp\left(-\frac{\tau^2}{2\sigma^2}\right) + \exp\left(-\frac{\alpha\tau}{2}\right),$$

where $\Gamma \geq \sum_{j=1}^{\log_2(2n\gamma)}\mathcal{N}_2(\mathcal{G},2^{-j}\gamma,\boldsymbol{z})^{-2}$ and $\sigma = 12\int_{\frac{1}{n}}^{\gamma}\sqrt{n\log\mathcal{N}_2(\mathcal{G},\delta,\boldsymbol{z})}d\delta$.

Observe that the probability of deviation has both subgaussian and subexponential components.

Using the above result and Proposition 11 leads to useful bounds on the quantities in Lemma 7 for specific types of adaptive rates. Given a tree $\boldsymbol{z}$, we obtain a bound on the expected size of the sequential Rademacher process when we subtract off the data-dependent ℓ_2-norm of the function on the tree $\boldsymbol{z}$, adjusted by logarithmic terms.

Corollary 6. Suppose $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$, and let $\boldsymbol{z}$ be any $\mathcal{Z}$-valued tree of depth n. Assume $\log\mathcal{N}_2(\mathcal{G},\delta,n) \leq \delta^{-p}$ for some $p < 2$. Then

$$\mathbb{E}\sup_{g\in\mathcal{G},\gamma}\Bigg\{\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 4\sqrt{2(\log n)\log\mathcal{N}_2(\mathcal{G},\gamma/2,\boldsymbol{z})\left(\sum_{t=1}^{n}g^2(\boldsymbol{z}_t(\epsilon))+1\right)}$$
$$-24\sqrt{2}\log n\int_{1/n}^{\gamma}\sqrt{n\log\mathcal{N}_2(\mathcal{G},\delta,\boldsymbol{z})}d\delta\Bigg\} \leq 7 + 2\log n\,.$$

The next corollary yields slightly faster rates than Corollary 6 when $|\mathcal{G}| < \infty$.

Corollary 7. Suppose $\mathcal{G} \subseteq [-1,1]^{\mathcal{Z}}$ with $|\mathcal{G}| = N$, and let $\boldsymbol{z}$ be any $\mathcal{Z}$-valued tree of depth n. Then

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left\{\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\log\left(\log N\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e\right)\sqrt{32\left(\log N\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e\right)}\right\} \leq 1.$$

6.4 Achievable Rates

In this section we use Lemma 7 along with the probabilistic tools from the previous section to obtain an array of achievable adaptive regret bounds for various online learning problems. We subdivide the section into subsections based on the categories for adaptive regret bounds described in Section 6.1.1.

6.4.1 Model Adaptation

In this subsection we focus on achievable rates for oracle inequalities and model selection, but without dependence on data. The form of the rate is therefore $\mathcal{B}(f)$. Assume we have a class $\mathcal{F} = \bigcup_{R\geq 1} \mathcal{F}(R)$, with the property that $\mathcal{F}(R) \subseteq \mathcal{F}(R')$ for any $R \leq R'$. Let us adopt the abbreviation $\mathcal{R}_n := \mathcal{R}_n^{\mathrm{seq}}$. If we are told by an oracle that regret will be measured with respect to those hypotheses $f \in \mathcal{F}$ with $R(f) := \inf\{R : f \in \mathcal{F}(R)\} \leq R^*$, then using the minimax algorithm one can guarantee a regret bound of at most the sequential Rademacher complexity $\mathcal{R}_n(\mathcal{F}(R^*))$. On the other hand, given the optimality of the sequential Rademacher complexity for online learning problems for commonly encountered losses, we can argue that for any $f \in \mathcal{F}$ chosen in hindsight, one cannot expect a regret better than order $\mathcal{R}_n(\mathcal{F}(R(f)))$. In this section we show that simultaneously for all $f \in \mathcal{F}$, one can attain an adaptive upper bound of $O\left(\mathcal{R}_n(\mathcal{F}(R(f)))\sqrt{\log\left(\mathcal{R}_n(\mathcal{F}(R(f)))\right)}\log^{3/2} n\right)$. That is, we may predict as if we knew the optimal radius, at the price of a logarithmic factor. This is the price of adaptation.

Corollary 8. For any class of predictors $\mathcal{F}$ with $\mathcal{F}(1)$ non-empty, if one considers the supervised learning problem with 1-Lipschitz loss ℓ, the following rate is achievable:

$$\mathcal{B}(f) = \log^{3/2} n\left(K_1\mathcal{R}_n(\mathcal{F}(2R(f)))\left(1+\sqrt{\log\left(\frac{\log(2R(f))\cdot\mathcal{R}_n(\mathcal{F}(2R(f)))}{\mathcal{R}_n(\mathcal{F}(1))}\right)}\right)+K_2\Gamma\mathcal{R}_n(\mathcal{F}(1))\right),$$

for absolute constants K_1, K_2, and Γ defined in Lemma 9.

In fact, this statement is true more generally with $\mathcal{F}(2R(f))$ replaced by $\ell \circ \mathcal{F}(2R(f))$. It is tempting to attempt to prove the above statement with the exponential weights algorithm running as an aggregation procedure over the solutions for each R. In general, this approach will fail for two reasons. First, if function values grow with R, the exponential weights bound will scale linearly with this value. Second, an experts bound yields only a slower $\sqrt{n}$ rate. In Chapter 9 we show that by combining the techniques of this chapter with those of Chapter 5, we can overcome these difficulties to derive universal and efficient algorithms for model selection in online convex optimization and online learning.

As a special case of the above lemma, we obtain an *online PAC-Bayesian theorem*. We postpone this example to the next sub-section where we get a *data-dependent* version of this result. We now provide a bound for online linear optimization in 2-smooth Banach spaces that automatically adapts to the norm of the comparator. To prove it, we use Pinelis' concentration bound (Lemma 8) within the proof of the above corollary to remove the extra logarithmic factors.

Example 8 (Unconstrained Linear Optimization). *Consider linear optimization with $\mathcal{Y}$ being the unit ball of some reflexive Banach space with norm $\|\cdot\|_*$. Let $\mathcal{F} = \widehat{\mathcal{Y}}$ be the dual space and the loss $\ell(\hat{y}, y) = \langle \hat{y}, y \rangle$ (where we are using $\langle \cdot, \cdot \rangle$ to represent the linear functional in the first argument to the second argument). Define $\mathcal{F}(R) = \{f \mid \|f\| \leq R\}$ where $\|\cdot\|$ is the norm dual to $\|\cdot\|_*$. If the unit ball of $\mathcal{Y}$ is $(2, D)$-smooth, then the following rate is achievable for all f with $\|f\| \geq 1$:*

$$\mathcal{B}(f) = D\sqrt{n}\Big(8\|f\|\Big(1 + \sqrt{\log(2\|f\|) + \log\log(2\|f\|)}\Big) + 12\Big).$$

For special case of Hilbert spaces, this bound was achieved by McMahan and Orabona (2014).

6.4.2 Adapting to Data and Model Simultaneously

We now study achievable bounds that perform online model selection in a data-adaptive way. Of particular note is a new online optimistic PAC-Bayesian bound. This bound should be compared to Luo and Schapire (2015) and Koolen and van Erven (2015), with the reader noting that it is independent of the number of experts, is algorithm-independent, and depends quadratically on the expected loss of the expert we compare against.

Example 9 (Generalized Predictable Sequences (Supervised Learning)). *Consider an online supervised learning problem with a convex* 1*-Lipschitz loss. Let* $(M_t)_{t\geq 1}$ *be any predictable sequence that the learner can compute at round* t *based on information provided so far, including* x_t *(One can think of the predictable sequence* M_t *as a prior guess for the hypothesis we would compare with in hindsight). Then the following adaptive rate is achievable:*

$$\mathcal{B}(f;x_{1:n}) = \inf_{\gamma}\Bigg\{ K_1\sqrt{\log n\cdot\log\mathcal{N}_2(\mathcal{F},\gamma/2,n)\cdot\left(\sum_{t=1}^{n}\left(f(x_t)-M_t\right)^2+1\right)} \\ +K_2\log n\int_{1/n}^{\gamma}\sqrt{n\log\mathcal{N}_2(\mathcal{F},\delta,n)}d\delta + 2\log n+7\Bigg\},$$

for constants $K_1 = 4\sqrt{2}, K_2 = 24\sqrt{2}$ *from Corollary 6. If we assume that the sequential covering of class* $\mathcal{F}$ *grows as* $\log\mathcal{N}_2(\mathcal{F},\varepsilon,n)\leq\varepsilon^{-p}$ *for some* $p<2$*, we get that*

$$\mathcal{B}(f) = \tilde{O}\left(\left(\sqrt{\sum_{t=1}^{n}\left(f(x_t)-M_t\right)^2+1}\right)^{1-\frac{p}{2}}\left(\sqrt{n}\right)^{p/2}\right).$$

As p *approaches zero, we get full adaptivity and are able to replace* n *by* $\sum_{t=1}^{n}\left(f(x_t)-M_t\right)^2+1$*. On the other hand, as* p *gets closer to* 2 *(i.e. more complex function classes), we do not adapt and get a uniform bound in terms of* n*. For* $p\in(0,2)$*, we attain a natural interpolation.*

The achievability of Example 9 is a direct consequence of Eq. (6.3) in Lemma 7, followed by Corollary 6 (one can include any predictable sequence in the Rademacher average part because $\sum_t M_t\epsilon_t$ is zero mean).

Example 10 (Regret to Fixed Vs Regret to Best (Supervised Learning)). *Consider an online supervised learning problem with a convex* 1*-Lipschitz loss and let* $|\mathcal{F}| = N$*. Let* $f^\star \in \mathcal{F}$ *be a fixed expert chosen in advance. The following bound is achievable:*

$$\mathcal{B}(f, x_{1:n}) = 4\log\left(\log N \sum_{t=1}^{n}(f(x_t) - f^\star(x_t))^2 + e\right)\sqrt{32\left(\log N \sum_{t=1}^{n}(f(x_t) - f^\star(x_t))^2 + e\right)} + 2.$$

In particular, against $f^\star$ *we have* $\mathcal{B}(f^\star, x_{1:n}) = O(1)$*, and against an arbitrary expert we have* $\mathcal{B}(f, x_{1:n}) = O\big(\sqrt{n\log N}(\log{(n \cdot \log N)})\big)$.

Example 10 follows from Eq. (6.3) in Lemma 7 followed by Corollary 7. The example extends the study of Even-Dar et al. (2008) to supervised learning and general class of experts $\mathcal{F}$.

Example 11 (Optimistic PAC-Bayes). *Assume that we have a countable set of experts and that the loss for each expert on any round is non-negative and bounded by* 1*. The function class* $\mathcal{F}$ *is the set of all distributions over these experts, and* $\mathcal{X} = \{0\}$*. This setting can be formulated as online linear optimization where the loss of mixture* f *over experts, given instance* y*, is* $\langle f, y\rangle$*, the expected loss under the mixture. The following adaptive bound is achievable:*

$$\mathcal{B}(f; y_{1:n}) = \sqrt{50\left(\mathrm{KL}(f|\pi) + \log(n)\right)\sum_{t=1}^{n}\mathop{\mathbb{E}}_{i\sim f}\left\langle e_i, y_t\right\rangle^2} + 50\left(\mathrm{KL}(f|\pi) + \log(n)\right) + 10.$$

This adaptive bound is an **online PAC-Bayesian bound**. The rate adapts not only to the KL divergence of f with fixed prior π but also replaces n with $\sum_{t=1}^{n}\mathbb{E}_{i\sim f}\left\langle e_i, y_t\right\rangle^2$. Note that we have $\sum_{t=1}^{n}\mathbb{E}_{i\sim f}\left\langle e_i, y_t\right\rangle^2 \le \sum_{t=1}^{n}\langle f, y_t\rangle$, yielding the small-loss type bound described earlier. This is an improvement over the bound in Luo and Schapire (2015) in that the bound is independent of number of experts, and so holds even for countably infinite sets of experts. The KL term in our bound may be compared to the MDL-style term in the bound of Koolen and van Erven (2015). If we have a large (but finite) number of experts and take π to be uniform, the above bound provides an improvement over both Chaudhuri et al. (2009)[2] and Luo and Schapire (2015).

Evaluating the above bound with a distribution f that places all its weight on any one expert appears to address the open question posed by Cesa-Bianchi et al. (2007) of obtaining algorithm-independent oracle-type variance bounds for experts. The proof of achievability of the above rate is deferred to Section 6.5 because it requires a slight variation on the symmetrization lemma specific to the problem.

6.5 Detailed Proofs

Proof of Lemma 7. We first prove equation (6.2). We start from the definition of $\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B})$. Our proof proceeds "inside out" by starting with the n^{th} term and then working backwards by repeatedly applying the minimax theorem (Section 2.6). We start with the innermost term as,

$$\begin{aligned}
&\sup_{x_n\in\mathcal{X}}\inf_{q_n\in\Delta(\widehat{\mathcal{Y}})}\sup_{y_n\in\mathcal{Y}}\left(\mathbb{E}_{\hat{y}_n\sim q_n}\left[\ell(\hat{y}_n,y_n)-\inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^n\ell(f(x_t),y_t)+\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right]\right)\\
&=\sup_{x_n\in\mathcal{X}}\inf_{q_n\in\Delta(\widehat{\mathcal{Y}})}\sup_{p_n\in\Delta(\mathcal{Y})}\left(\mathbb{E}_{\substack{\hat{y}_n\sim q_n\\ y_n\sim p_n}}\left[\sum_{t=1}^n\ell(\hat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^n\ell(f(x_t),y_t)+\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right]\right)\\
&=\sup_{x_n\in\mathcal{X}}\sup_{p_n\in\Delta(\mathcal{Y})}\inf_{q_n\in\Delta(\widehat{\mathcal{Y}})}\left(\mathbb{E}_{\substack{\hat{y}_n\sim q_n\\ y_n\sim p_n}}\left[\sum_{t=1}^n\ell(\hat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^n\ell(f(x_t),y_t)+\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right]\right)\\
&=\sup_{x_n\in\mathcal{X}}\sup_{p_n\in\Delta(\mathcal{Y})}\inf_{\hat{y}_n\in\widehat{\mathcal{Y}}}\left(\mathbb{E}_{y_n\sim p_n}\left[\sum_{t=1}^n\ell(\hat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^n\ell(f(x_t),y_t)+\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right]\right)\\
&=\sup_{x_n\in\mathcal{X}}\sup_{p_n\in\Delta(\mathcal{Y})}\left(\mathbb{E}_{y_n\sim p_n}\left[\sup_{f\in\mathcal{F}}\left\{\inf_{\hat{y}_n\in\widehat{\mathcal{Y}}}\mathbb{E}_{y_n\sim p_n}\left[\sum_{t=1}^n\ell(\hat{y}_t,y_t)\right]-\sum_{t=1}^n\ell(f(x_t),y_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right]\right).
\end{aligned}$$

To apply the minimax theorem in step 3 above, we note that the term in the round bracket is linear in q_n and in p_n (as it is an expectation). Hence under mild assumptions on the sets $\widehat{\mathcal{Y}}$ and $\mathcal{Y}$, the losses, and the adaptive rate $\mathcal{B}$, one can apply a generalized version of the minimax theorem to swap $\sup_{p_n}$ and $\inf_{q_n}$. Compactness of the sets and lower semi-continuity of the losses and $\mathcal{B}$ are sufficient; see Section 2.6 for discussion. but see Rakhlin et al. (2015); Rakhlin and Sridharan (2012)

[2]See Luo and Schapire (2015) for a detailed discussion of the differences between KL-based bounds and quantile bounds.

for milder conditions. Proceeding backward from n to 1 in a similar fashion we end up with the following quantity:

$$
\begin{aligned}
&\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}, \mathcal{B}) \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\widehat{\mathcal{Y}})} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \ell(f(x_t), y_t) + \mathcal{B}(f; x_{1:n}, y_{1:n}) \right\} \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \inf_{\hat{y}_t \in \widehat{\mathcal{Y}}} \mathbb{E}_{y_t \sim p_t} \left[\ell(\hat{y}_t, y_t) \right] - \sum_{t=1}^{n} \ell(f(x_t), y_t) - \mathcal{B}(f; x_{1:n}, y_{1:n}) \right\} \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \mathbb{E}_{y_t' \sim p_t} \left[\ell(f(x_t), y_t') \right] - \ell(f(x_t), y_t) - \mathcal{B}(f; x_{1:n}, y_{1:n}) \right\} \right]. \quad (6.6)
\end{aligned}
$$

See Rakhlin and Sridharan (2012) for more details of the steps involved in obtaining the above equality.

To proceed, we use Jensen's inequality to pull out the expectations with respect to y_t''s, which gives

$$
\begin{aligned}
&\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \ell(f(x_t), y_t') - \ell(f(x_t), y_t) - \mathcal{B}(f; x_{1:n}, y_{1:n}) \right\} \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \sup_{y_t'' \in \mathcal{Y}} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \ell(f(x_t), y_t') - \ell(f(x_t), y_t) - \mathcal{B}(f; x_{1:n}, y_{1:n}'') \right\} \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t, y_t' \sim p_t} \mathop{\mathbb{E}}_{\epsilon_t} \sup_{y_t'' \in \mathcal{Y}} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t \left(\ell(f(x_t), y_t') - \ell(f(x_t), y_t) \right) - \mathcal{B}(f; x_{1:n}, y_{1:n}'') \right\} \right].
\end{aligned}
$$

We now move to an upper bound by allowing a supremum over y_t and y_t' at each step.

$$
\begin{aligned}
&\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{y_t, y_t' \in \mathcal{Y}} \mathop{\mathbb{E}}_{\epsilon_t} \sup_{y_t'' \in \mathcal{Y}} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t \left(\ell(f(x_t), y_t') - \ell(f(x_t), y_t) \right) - \mathcal{B}(f; x_{1:n}, y_{1:n}'') \right\} \right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\epsilon_t} \sup_{y_t'' \in \mathcal{Y}} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} 2\epsilon_t \ell(f(x_t), y_t) - \mathcal{B}(f; x_{1:n}, y_{1:n}'') \right\} \right] \\
&= \sup_{\mathbf{x}, \mathbf{y}, \mathbf{y}'} \mathbb{E}_{\epsilon} \left[\sup_{f \in \mathcal{F}} \left\{ 2 \sum_{t=1}^{n} \epsilon_t \ell(f(\mathbf{x}_t(\epsilon)), \mathbf{y}_t(\epsilon)) - \mathcal{B}(f; \mathbf{x}_{1:n}(\epsilon), \mathbf{y}_{2:n+1}'(\epsilon)) \right\} \right],
\end{aligned}
$$

where in the last step we switch to tree notation, but keep in mind that each y_t'' is picked after drawing ϵ_t, and thus the tree $\mathbf{y}'$ appears with one index shifted. This proves (6.2).

Finally, we proceed to prove inequality (6.3). Here, we employ the convexity assumption $\ell(\hat{y}_t, y_t) - \ell(f(x_t), y_t) \leq \ell'(\hat{y}_t, y_t)(\hat{y}_t - f(x_t))$, where the derivative is with respect to the first argument. As before, applying the minimax theorem,

$$\begin{aligned}
\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B}) &= \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \inf_{q_t\in\Delta(\widehat{\mathcal{Y}})} \sup_{y_t\in\mathcal{Y}} \mathop{\mathbb{E}}_{\hat{y}_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\ell(f(x_t),y_t)+\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&= \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \inf_{\hat{y}_t\in\widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n}\ell(\hat{y}_t,y_t) - \inf_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\ell(f(x_t),y_t)+\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&\le \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \inf_{\hat{y}_t\in\widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\ell'(\hat{y}_t,y_t)(\hat{y}_t-f(x_t))-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right].
\end{aligned}$$

We may now pick $\hat{y}_t=\hat{y}_t^*(p_t):=\arg\min_{\hat{y}}\mathbb{E}_{y_t\sim p_t}\left[\ell(\hat{y}_t,y_t)\right]$. By convexity (and assuming the loss allows swapping of derivative and expectation), $\mathbb{E}_{y_t\sim p_t}\left[\ell'(\hat{y}_t,y_t)\right]=0$. This (sub)optimal strategy yields an upper bound of

$$\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\left(\ell'(\hat{y}_t^*,y_t)-\mathbb{E}_{y_t'\sim p_t}\left[\ell'(\hat{y}_t^*,y_t')\right]\right)(\hat{y}_t^*-f(x_t))-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right].$$

Since $\left(\ell'(\hat{y}_t^*,y_t)-\mathbb{E}_{y_t'\sim p_t}\left[\ell'(\hat{y}_t^*,y_t')\right]\right)\hat{y}_t^*$ is independent of f and has expected value of 0, the above quantity is equal to

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\left(\mathbb{E}_{y_t'\sim p_t}\left[\ell'(\hat{y}_t^*,y_t')\right]-\ell'(\hat{y}_t^*,y_t)\right)f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&\le \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t,y_t'\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\left(\ell'(\hat{y}_t^*,y_t')-\ell'(\hat{y}_t^*,y_t)\right)f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&= \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t,y_t'\sim p_t} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}\epsilon_t\left(\ell'(\hat{y}_t^*,y_t')-\ell'(\hat{y}_t^*,y_t)\right)f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right].
\end{aligned}$$

Replacing $(\ell'(\hat{y}_t^*,y_t')-\ell'(\hat{y}_t^*,y_t))$ by $2Ls_t$ for $s_t\in[-1,1]$ and taking supremum over s_t we get,

$$\begin{aligned}
&\le \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{p_t\in\Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t,y_t'\sim p_t} \sup_{s_t\in[-1,1]} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}2L\epsilon_t s_t f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&\le \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{y_t} \sup_{s_t\in[-1,1]} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}2L\epsilon_t s_t f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right].
\end{aligned}$$

Since the suprema over s_t are achieved at $\{\pm 1\}$ by convexity, the last expression is equal to

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{y_t} \sup_{s_t\in\{-1,1\}} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}2L\epsilon_t s_t f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&= \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{y_t} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}2L\epsilon_t f(x_t)-\mathcal{B}(f;x_{1:n},y_{1:n})\right\}\right] \\
&= \sup_{\mathbf{x},\mathbf{y}}\mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{t=1}^{n}2L\epsilon_t f(\mathbf{x}_t(\epsilon))-\mathcal{B}(f;\mathbf{x}_{1:n}(\epsilon),\mathbf{y}_{1:n}(\epsilon))\right\}\right].
\end{aligned}$$

In the last but one step we removed s_t, since for any function Ψ, and any $s \in \{\pm 1\}$, $\mathbb{E}[\Psi(s\epsilon)] = \frac{1}{2}(\Psi(s) + \Psi(-s)) = \frac{1}{2}(\Psi(1) + \Psi(-1)) = \mathbb{E}[\Psi(\epsilon)]$. $\square$

Proof of Proposition 11. Define $Z_i = [X_i - B_i\theta_i]_+$. As long as $\theta_i \geq 1$, for any strictly positive τ we have the tail behavior

$$\mathbb{P}(Z_i \geq t) = \mathbb{P}(X_i - B_i\theta_i \geq \tau) \leq C_1 \exp\left(-\frac{(B_i(\theta_i - 1) + \tau)^2}{2\sigma_i^2}\right) + C_2 \exp\left(-(B_i(\theta_i - 1) + \tau)s_i\right).$$

Note that for any positive sequence $(\delta_i)_{i\in I}$ with $\delta = \sum_{i\in I} \delta_i$,

$$\mathbb{E}\left[\sup_{i\in I}\{X_i - B_i\theta_i\}\right] \leq \mathbb{E}\left[\sup_{i\in I} Z_i\right] \leq \sum_{i\in I}\mathbb{E}[Z_i] \leq \delta + \sum_{i\in I}\int_{\delta_i}^{\infty}\mathbb{P}(Z_i \geq \tau)d\tau.$$

The sum of the integrals above is equal to

$$\begin{aligned}
&\sum_{i\in I}\int_{\delta_i}^{\infty}\mathbb{P}(X_i - B_i\theta_i \geq \tau)d\tau \\
&\quad\leq C_1\sum_{i\in I}\int_0^{\infty}\exp\left(-\frac{(B_i(\theta_i-1)+\tau)^2}{2\sigma_i^2}\right)dt + C_2\sum_{i\in I}\int_0^{\infty}\exp\left(-\left(B_i(\theta_i-1)+\tau\right)s_i\right)d\tau \\
&\quad\leq C_1\sum_{i\in I}\exp\left(-\frac{1}{2}\left(\frac{B_i}{\sigma_i}\right)^2(\theta_i-1)^2\right)\int_0^{\infty}e^{-\frac{\tau^2}{2\sigma_i^2}}d\tau + C_2\sum_{i\in I}\exp\left(-B_is_i(\theta_i-1)\right)\int_0^{\infty}e^{-\tau s_i}d\tau \\
&\quad\leq \sqrt{\frac{\pi}{2}}C_1\sum_{i\in I}\sigma_i\exp\left(-\frac{1}{2}\left(\frac{B_i}{\sigma_i}\right)^2(\theta_i-1)^2\right) + C_2\sum_{i\in I}s_i^{-1}\exp\left(-B_is_i(\theta_i-1)\right) \\
&\quad\leq \frac{\pi^2\sqrt{\pi}}{6\sqrt{2}}C_1\bar{\sigma} + \frac{\pi^2}{6}C_2(\bar{s})^{-1},
\end{aligned}$$

where the last step is obtained by plugging in the expression

$$\theta_i = \max\left\{\frac{\sigma_i}{B_i}\sqrt{2\log(\sigma_i/\bar{\sigma}) + 4\log(i)}, (B_is_i)^{-1}\log\left(i^2(\bar{s}/s_i)\right)\right\} + 1$$

and using as an upper bound $\frac{\sigma_i}{B_i}\sqrt{2\log(i^2\sigma_i/\bar{\sigma})} + 1$ for θ_i in the sub-gaussian part and $(B_is_i)^{-1}\log\left(i^2\bar{s}/s_i\right) + 1$ for θ_i in the sub-exponential part. Since δ can be chosen arbitrarily small, we may over-bound the above constant and obtain the result. $\square$

Proof of Lemma 10. Fix $\gamma > 0$. For $j \geq 0$, let V_j be a minimal sequential cover of $\mathcal{G}$ on $\boldsymbol{z}$ at scale $\beta_j = 2^{-j}\gamma$ and with respect to empirical ℓ_2 norm. Let $\boldsymbol{v}^j[g, \epsilon]$ be an element guaranteed to be

β_j-close to f at the j-th level, for the given ϵ. Choose $N = \log_2(2\gamma n)$, so that $\beta_N n \le 1$. Let us use the shorthand $\mathcal{N}_2(\gamma) := \mathcal{N}_2(\mathcal{G}, \gamma, \boldsymbol{z})$.

For any $\epsilon \in \{\pm 1\}^n$ and $g \in \mathcal{G}$,

$$\sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\alpha g(\boldsymbol{z}_t(\epsilon))^2$$

can be written as

$$\begin{aligned}
&\sum_{t=1}^{n} \left(\epsilon_t(g(\boldsymbol{z}_t(\epsilon)) - \boldsymbol{v}_t^0[g,\epsilon](\epsilon))\right) + \sum_{t=1}^{n} \left(\epsilon_t \boldsymbol{v}_t^0[g,\epsilon](\epsilon) - 2\alpha g(\boldsymbol{z}_t(\epsilon))^2\right) \\
&\le \sum_{t=1}^{n} \left(\epsilon_t(g(\boldsymbol{z}_t(\epsilon)) - \boldsymbol{v}_t^0[g,\epsilon](\epsilon))\right) + \sum_{t=1}^{n} \left(\epsilon_t \boldsymbol{v}_t^0[g,\epsilon](\epsilon) - \alpha \boldsymbol{v}_t^0[g,\epsilon](\epsilon)^2\right) \\
&= \sum_{t=1}^{n} \left(\epsilon_t(g(\boldsymbol{z}_t(\epsilon)) - \boldsymbol{v}_t^N[g,\epsilon](\epsilon)\right) + \sum_{t=1}^{n}\sum_{k=1}^{N} \epsilon_t \left(\boldsymbol{v}_t^k[g,\epsilon](\epsilon) - \boldsymbol{v}_t^{k-1}[g,\epsilon](\epsilon)\right) \\
&\qquad + \sum_{t=1}^{n} \left(\epsilon_t \boldsymbol{v}_t^0[g,\epsilon](\epsilon) - \alpha \boldsymbol{v}_t^0[g,\epsilon](\epsilon)^2\right).
\end{aligned}$$

By Cauchy-Schwartz, the first term is upper bounded by $n\beta_N \le 1$. The second term above is upper bounded by

$$\sum_{k=1}^{N}\sum_{t=1}^{n} \epsilon_t \left(\boldsymbol{v}_t^k[g,\epsilon](\epsilon) - \boldsymbol{v}_t^{k-1}[g,\epsilon](\epsilon)\right) \le \sum_{k=1}^{N} \sup_{\boldsymbol{w}^k \in W_k} \sum_{t=1}^{n} \epsilon_t \boldsymbol{w}_t^k(\epsilon),$$

where W_k is a set of differences of trees for levels k and $k-1$ (see (Rakhlin et al., 2015, Proof of Theorem 3)). Finally, the third term is controlled by

$$\sum_{t=1}^{n} \left(\epsilon_t \boldsymbol{v}_t^0[g,\epsilon](\epsilon) - \alpha \boldsymbol{v}_t^0[g,\epsilon](\epsilon)^2\right) \le \sup_{\boldsymbol{v} \in V_0} \sum_{t=1}^{n} \left(\epsilon_t \boldsymbol{v}_t(\epsilon) - \alpha \boldsymbol{v}_t^2(\epsilon)\right).$$

The probability expression in the statement of the lemma can now be upper bounded by

$$\begin{aligned}
\mathbb{P}\Bigg(&\sum_{k=1}^{N} \sup_{\boldsymbol{w}^k \in W_k} \sum_{t=1}^{n} \epsilon_t \boldsymbol{w}_t^k(\epsilon) + \sup_{\boldsymbol{v} \in V_0} \sum_{t=1}^{n} \left(\epsilon_t \boldsymbol{v}_t(\epsilon) - \alpha \boldsymbol{v}_t^2(\epsilon)\right) \\
&\qquad - \frac{\log \mathcal{N}_2(\gamma)}{\alpha} - c\int_{1/n}^{\gamma} \sqrt{n \log \mathcal{N}_2(\delta)}\, d\delta > \tau\Bigg).
\end{aligned}$$

In view of the inequality

$$\sqrt{72}\sum_{k=1}^{N} \beta_k \sqrt{n \log \mathcal{N}_2(\beta_k)} \le 12\sqrt{2}\int_{1/n}^{\gamma} \sqrt{n \log \mathcal{N}_2(\delta)}\, d\delta,$$

this probability can be further upper bounded by

$$\mathbb{P}\left(\sum_{k=1}^{N}\sup_{\boldsymbol{w}^k\in W_k}\sum_{t=1}^{n}\epsilon_t\boldsymbol{w}_t^k(\epsilon)+\sup_{\boldsymbol{v}\in V_0}\sum_{t=1}^{n}\left(\epsilon_t\boldsymbol{v}_t(\epsilon)-\alpha\boldsymbol{v}_t^2(\epsilon)\right)\right.$$
$$\left.-\frac{\log\mathcal{N}_2(\gamma)}{\alpha}-\sqrt{72}\sum_{k=1}^{N}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}>\tau\right).$$

Define a distribution p on $\{1,\ldots,N\}$ by $p_k=\frac{\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}}{\sum_{k=1}^{N}\beta_j\sqrt{n\log\mathcal{N}_2(\beta_j)}}$. Then the probability above can be upper bounded by

$$\mathbb{P}\left(\exists k\in[N]\text{ s.t. }\sup_{\boldsymbol{w}^k\in W_k}\sum_{t=1}^{n}\epsilon_t\boldsymbol{w}_t^k(\epsilon)-\sqrt{72}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}>\frac{\tau p_k}{2}\right.$$
$$\left.\vee\sup_{\boldsymbol{v}\in V_0}\sum_{t=1}^{n}\left(\epsilon_t\boldsymbol{v}_t(\epsilon)-\alpha\boldsymbol{v}_t^2(\epsilon)\right)-\frac{\log\mathcal{N}_2(\gamma)}{\alpha}>\frac{\tau}{2}\right)$$
$$\leq\sum_{k=1}^{N}\mathbb{P}\left(\sup_{\boldsymbol{w}^k\in W_k}\sum_{t=1}^{n}\epsilon_t\boldsymbol{w}_t^k(\epsilon)-\sqrt{72}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}>\frac{\tau p_k}{2}\right)$$
$$+\mathbb{P}\left(\sup_{\boldsymbol{v}\in V_0}\sum_{t=1}^{n}\left(\epsilon_t\boldsymbol{v}_t(\epsilon)-\alpha\boldsymbol{v}_t^2(\epsilon)\right)-\frac{\log\mathcal{N}_2(\gamma)}{\alpha}>\frac{\tau}{2}\right).$$

The second term can be upper bounded using Chernoff method by

$$\sum_{\boldsymbol{v}\in V_0}\mathbb{P}\left(\sum_{t=1}^{n}\left(\epsilon_t\boldsymbol{v}_t(\epsilon)-\alpha\boldsymbol{v}_t^2(\epsilon)\right)-\frac{\log\mathcal{N}_2(\gamma)}{\alpha}>\frac{\tau}{2}\right)$$
$$\leq\mathcal{N}_2(\gamma)\exp\left(-\frac{\alpha\tau}{2}-\log\mathcal{N}_2(\gamma)\right)\leq\exp\left(-\frac{\alpha\tau}{2}\right),$$

while the first sum of probabilities can be upper bounded by

$$\sum_{k=1}^{N}\sum_{\boldsymbol{w}^k\in W_k}\mathbb{P}\left(\sum_{t=1}^{n}\epsilon_t\boldsymbol{w}_t^k(\epsilon)-\sqrt{72}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}>\frac{\tau\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}}{2\sum_{k=1}^{N}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}}\right).\tag{6.7}$$

For any k, the tail probability above is controlled by Hoeffding-Azuma inequality as

$$\mathbb{P}\left(\sum_{t=1}^{n}\epsilon_t\boldsymbol{w}_t^k(\epsilon)>\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}\left(6\sqrt{2}+\frac{\tau}{2\sum_{k=1}^{N}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}}\right)^2\right)$$
$$\leq\exp\left(-\frac{1}{18}\log\mathcal{N}_2(\beta_k)\left(6\sqrt{2}+\frac{\tau}{2\sum_{k=1}^{N}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}}\right)^2\right)$$
$$\leq\exp\left(-4\log\mathcal{N}_2(\beta_k)\right)\exp\left(-\frac{\tau^2}{18\left(2\sum_{k=1}^{N}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}\right)^2}\right),$$

because $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{w}_t^k(\epsilon)^2 \le 3\beta_k^2$ for any ϵ by triangle inequality (see Rakhlin et al. (2015)). Then the double sum in (6.7) is upper bounded by

$$\Gamma \exp\left(-\frac{\tau^2}{18\left(2\sum_{k=1}^{N}\beta_k\sqrt{n\log\mathcal{N}_2(\beta_k)}\right)^2}\right),$$

where $\Gamma \ge \sum_{k=1}^{N}\mathcal{N}_2(\beta_k)^{-2}$. This upper bound can be further relaxed to

$$\Gamma \exp\left(-\frac{\tau^2}{2\left(12\int_{1/n}^{\gamma}\sqrt{n\log\mathcal{N}_2(\delta)}d\delta\right)^2}\right).$$

Since $N = \log_2(2\gamma n)$, we may take

$$\Gamma = \sum_{k=1}^{\log_2(2\gamma n)} \mathcal{N}_2(\gamma 2^{-k})^{-2}.$$

$\square$

Proof of Corollary 6. Let $\mathcal{N}_2(\gamma) := \mathcal{N}_2(\mathcal{G},\gamma,\boldsymbol{z})$. Observe that

$$2\sqrt{2\log n\log\mathcal{N}_2(\gamma/2)\left(\sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon))+1\right)} = \inf_{\alpha>0}\left\{\frac{\log n\log\mathcal{N}_2(\gamma/2)}{\alpha} + 2\alpha\left(\sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon))+1\right)\right\}.$$

Furthermore, the optimal value of α is

$$\sqrt{\frac{(\log n)\left(\log\mathcal{N}_2(\gamma/2)\right)}{2(\sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon))+1)}},$$

which is a number between $d_\ell = \sqrt{\frac{(\log n)(\log\mathcal{N}_2(\gamma/2))}{2(n+1)}}$ and $d_u = \sqrt{(\log n)\left(\log\mathcal{N}_2(\gamma/2)\right)}$ as long as $\mathcal{N}_2(\gamma/2) > 1$. With this we get

$$\begin{aligned}
&\sup_{\substack{g\in\mathcal{G}\\ \gamma\in[n^{-1},1]}}\Bigg[\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 4\sqrt{2(\log n)\left(\log\mathcal{N}_2(\gamma/2)\right)\left(\sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon))+1\right)} \\
&\qquad\qquad -24\sqrt{2}\log n\int_{1/n}^{\gamma}\sqrt{n\log\mathcal{N}_2(\delta)}d\delta + 2\log n\Bigg] \\
&\le \sup_{\substack{g\in\mathcal{G}\\ \gamma\in[n^{-1},1],\alpha\in[d_\ell,d_u]}}\Bigg[\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - \frac{2(\log n)\left(\log\mathcal{N}_2(\gamma/2)\right)}{\alpha} - 4\alpha\sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon)) \\
&\qquad\qquad -24\sqrt{2}\log n\int_{1/n}^{\gamma}\sqrt{n\log\mathcal{N}_2(\delta)}d\delta - 2\log n\Bigg].
\end{aligned} \tag{6.8}$$

The case of $\gamma \in [1/n, 2/n)$ will be considered separately. Let us assume $\gamma \geq 2/n$. We now discretize both α and γ by defining $\alpha_i = 2^{-(i-1)}d_u$ and $\gamma_j = 2^j n^{-1}$, $i, j \geq 1$. We go to an upper bound by mapping each α to α_i or $\alpha_i/2$, depending on the direction of the sign. Similarly, we map γ to either γ_i or $2\gamma_i$. The upper bound becomes

$$\max_{i,j} \sup_{g\in\mathcal{G}} \sum_{t=1}^{n} \left(\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\alpha_i g^2(\boldsymbol{z}_t(\epsilon))\right) - (2\log n)\left(\frac{\log \mathcal{N}_2(\gamma_j)}{\alpha_i} + 12\sqrt{2}\int_{1/n}^{\gamma_j} \sqrt{n \log \mathcal{N}_2(\delta)}d\delta + 1\right).$$

Given the doubling nature of α_i and γ_j, the indices i, j are upper bounded by $O(\log n)$. Now define a collection of random variables indexed by (i, j)

$$X_{i,j} = \sup_{g\in\mathcal{G}} \sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\alpha_i g^2(\boldsymbol{z}_t(\epsilon))$$

and constants

$$B_{i,j} = \frac{\log \mathcal{N}_2(\gamma_j)}{\alpha_i} + 12\sqrt{2}\int_{1/n}^{\gamma_j} \sqrt{n \log \mathcal{N}_2(\delta)}d\delta + 1.$$

Lemma 10 establishes that

$$\mathbb{P}(X_{i,j} - B_{i,j} > \tau) \leq \Gamma \exp\left(-\frac{\tau^2}{2\sigma_j^2}\right) + \exp\left(-\frac{\alpha_i \tau}{2}\right)$$

where $\sigma_j = 12\sqrt{2}\int_{\frac{1}{n}}^{\gamma_j} \sqrt{n \log \mathcal{N}_2(\delta)}d\delta$ and Γ as specified in Lemma 10. Whenever δ-entropy grows as δ^{-p}, $\sigma_j \leq 12\sqrt{2}\sqrt{n}$, ensuring $\log(\sigma_j/\sigma_1) \leq \log(n)$. Further, we can take $1 \leq \Gamma \leq \log(2n)$.

Proposition 11 is used with a sequence of random variables, but we can easily put the pairs (i, j) into a vector of size at most $\log_2(n)^2$. Observe that $s_i = \alpha_i/2$, $(B_{i,j}s_i)^{-1} \leq 2$, $\sigma_j/B_{i,j} \leq 1$, $s_1/s_i \leq \sqrt{2(n+1)}$. Then, by taking $\bar{\sigma} = \min\{1/\Gamma, \sigma_1\}$ and $\bar{s} = s_1$,

$$\begin{aligned}
\theta_{k_{i,j}} &= \max\left\{\frac{\sigma_j}{B_{i,j}}\sqrt{2\log(\sigma_j/\bar{\sigma}) + 4\log(k_{i,j})}, (B_{i,j}s_i)^{-1}\log\left(k_{i,j}^2(\bar{s}/s_i)\right)\right\} + 1 \\
&\leq \max\left\{\sqrt{2\log(n) + 2\log(\log(2n)) + 4\log(k_{i,j})}, 2\log\left(k_{i,j}^2\sqrt{2(n+1)}\right)\right\} + 1
\end{aligned}$$

where $k_{i,j} = (\log n) \cdot (i-1) + j$. This choice of the multiplier ensures

$$\mathbb{E}\max_{i,j}\left\{X_{i,j} - \theta_{k_{i,j}}B_{i,j}\right\} \leq 3\Gamma\bar{\sigma} + 4\alpha_1^{-1} \leq 7$$

and $\theta_{i,j}$ is shown to be upper bounded by $2\log n$. Hence,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G},\gamma}\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 4\sqrt{2\log n\log\mathcal{N}_2(\gamma/2)\left(\sum_{t=1}^{n}g^2(\boldsymbol{z}_t(\epsilon))+1\right)}\right.$$
$$\left. -24\sqrt{2}\log n\int_{\frac{1}{n}}^{\gamma}\sqrt{n\log\mathcal{N}_2(\delta)}d\delta\right]$$
$$\le 7+2\log n.$$

Now, consider the case $\gamma\in[1/n,2/n)$. We upper bound (6.8) by

$$\max_i\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\left(\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\alpha_i g^2(\boldsymbol{z}_t(\epsilon))\right) - (2\log n)\left(\frac{\log\mathcal{N}_2(1/n)}{\alpha_i}+1\right),$$

which is controlled by setting $\gamma=1/n$ in Lemma 10. This case is completed by invoking Proposition 11 as before. □

Proof of Corollary 7. Assume $N>e$ and let $C>0$. We first note that

$$\inf_{\alpha>0}\left\{\frac{C\log\left(\frac{\sqrt{C}\log N}{\alpha}\right)\log N}{\alpha}+\alpha\left(\sum_{t=1}^{n}g^2(\boldsymbol{z}_t(\epsilon))+\frac{e}{\log N}\right)\right\}$$
$$\le 2\log\left(\log N\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e\right)\sqrt{C\left(\log N\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e\right)},$$

with the inequality obtained using $\alpha^\star=\sqrt{\frac{C\log N}{\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e/\log N}}$, which is a number between $d_\ell := \sqrt{\frac{C\log N}{n+e/logN}}$ and $d_u := \sqrt{\frac{C}{e}}\log N$. Consequently,

$$\sup_{g\in\mathcal{G}}\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - 2\log\left(\log N\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e\right)\sqrt{C\left(\log N\sum_{t=1}^{n}g^2(\boldsymbol{z}(\epsilon))+e\right)}$$
$$\le \sup_{\substack{g\in\mathcal{G}\\ \alpha\in[d_\ell,d_u]}}\left[\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - \alpha\sum_{t=1}^{n}g^2(\boldsymbol{z}_t(\epsilon)) - \frac{C\log N}{\alpha}\log\left(\frac{\sqrt{C}\log N}{\alpha}\right)\right].$$

Let $L=\left\lceil\log_2\left(\sqrt{\frac{n\log N}{e}+1}\right)+1\right\rceil$. We discretize the range of α by defining $\alpha_i = d_u 2^{-(i-1)}$ for $i\in[L]$. The following upper bound holds:

$$\sup_{\substack{g\in\mathcal{G}\\ i\in[L]}}\left[\sum_{t=1}^{n}\epsilon_t g(\boldsymbol{z}_t(\epsilon)) - \frac{\alpha_i}{2}\sum_{t=1}^{n}g^2(\boldsymbol{z}_t(\epsilon)) - \frac{C\log N}{\alpha_i}\log\left(\frac{\sqrt{C}\log N}{\alpha_i}\right)\right].$$

Define a collection of random variables indexed by $i \in [L]$ with

$$X_i = \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) - \frac{\alpha_i}{2} \sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon)) \right]$$

and let $B_i = \frac{4 \log N}{\alpha_i}$. Applying Lemma 10 with $\gamma = 1/n$ establishes

$$\mathbb{P}(X_i - B_i > \tau) \leq \exp\left(-\frac{\alpha_i \tau}{8}\right).$$

We now set $s_i = \alpha_i/8$ and $\bar{s} = s_1$, and apply Proposition 11, yielding

$$\mathbb{E}\{X_i - B_i \theta_i\} \leq \frac{16\sqrt{e}}{C}.$$

It remains to relate this quantity to the rate we are trying to achieve. Note that our bound on $\mathbb{P}(X_i - B_i > \tau)$ has a pure exponential tail, so we only need to consider $\theta_i = (B_i s_i)^{-1} \log(i^2(\bar{s}/s_i)) + 1$. Taking $C \geq 32$ and observing that $(B_i s_i)^{-1} \leq 2$, we obtain

$$\begin{aligned} \theta_i &= (B_i s_i)^{-1} \log(i^2(\bar{s}/s_i)) + 1 \leq 2\log(i^2(\bar{s}/s_i)) + 1 = 2\log(i^2 2^{i-1}) + 1 \leq 2\log\left(i^2 2^i\right) \\ &\leq \frac{C}{4} \log\left(\frac{\sqrt{C}\log N}{\alpha_i}\right). \end{aligned}$$

Finally, we have

$$\sup_{\substack{g \in \mathcal{G} \\ i \in [L]}} \left[\sum_{t=1}^{n} \epsilon_t g(\boldsymbol{z}_t(\epsilon)) - \frac{\alpha_i}{2} \sum_{t=1}^{n} g^2(\boldsymbol{z}_t(\epsilon)) - \frac{32 \log N}{\alpha_i} \log\left(\frac{\sqrt{32}\log N}{\alpha_i}\right) \right] \leq \mathbb{E}\{X_i - B_i \theta_i\} \leq \frac{\sqrt{e}}{2} \leq 1.$$

□

Proof of Corollary 8. We prove the corollary for the convex Lipschitz loss setting from (6.3).

Our starting point to proving the bounds is Lemma 7, equation (6.2). To show achievability it suffices to show that

$$\begin{aligned} &\mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) - K_1 \mathcal{R}_n(\mathcal{F}(2R(f))) \log^{3/2} n \left(1 + \sqrt{\log\left(\frac{\mathcal{R}_n(\mathcal{F}(2R(f)))}{\mathcal{R}_n(\mathcal{F}(R(1)))}\right) + \log(\log(2R(f)))}\right) \\ &\qquad \leq K_2 \Gamma \mathcal{R}_n(\mathcal{F}(1)) \log^{3/2} n \end{aligned}$$

where Γ is the constant that will be inherited from Lemma 9. Define $R_i = 2^i$ and note that since the Rademacher complexity of the class $\mathcal{F}(R)$ is non-decreasing with R,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) - K_1 \mathcal{R}_n(\mathcal{F}(2R(f))) \log^{3/2} n \left(1 + \sqrt{\log\left(\frac{\mathcal{R}_n(\mathcal{F}(2R(f)))}{\mathcal{R}_n(\mathcal{F}(1))}\right) + \log(\log(2R(f)))}\right) \\
&= \sup_{R\geq 1} \sup_{f\in\mathcal{F}(R)} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) - K_1 \mathcal{R}_n(\mathcal{F}(2R)) \log^{3/2} n \left(1 + \sqrt{\log\left(\frac{\mathcal{R}_n(\mathcal{F}(2R))}{\mathcal{R}_n(\mathcal{F}(1))}\right) + \log(\log(2R))}\right) \\
&\leq \max_{i\in\mathbb{N}} \sup_{f\in\mathcal{F}(R_i)} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) - K_1 \mathcal{R}_n(\mathcal{F}(R_i)) \log^{3/2} n \left(1 + \sqrt{\log\left(\frac{\mathcal{R}_n(\mathcal{F}(R_i))}{\mathcal{R}_n(\mathcal{F}(1))}\right) + \log(\log(R_i))}\right).
\end{aligned} \tag{6.9}$$

Denote a shorthand $C_n = \sqrt{96\log^3(en^2)}$ and $D_n^i = \mathcal{R}_n(\mathcal{F}(R_i))$. Now note that by Lemma 9 we have that for every i and every $\theta > 1$,

$$\mathbb{P}_\epsilon\left(\sup_{f\in\mathcal{F}(R_i)}\left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\epsilon))\right| > 8\left(1+\theta C_n\right)\cdot D_n^i\right) \leq 2\Gamma e^{-3\theta^2} .$$

Let $X_i = \sup_{f\in\mathcal{F}(R_i)} \left|\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\epsilon))\right|$ and let $B_i = 8\left(1+C_n\right)\cdot D_n^i$. In this case rewriting the above one sided tail bound appropriately (with $\theta = 1 + \tau/(8C_n D_n^i)$) we see that for any $\tau > 0$,

$$\mathbb{P}(X_i - B_i > \tau) \leq \frac{2\Gamma}{e^3}\exp\left(-\frac{\tau^2}{2^8\log^3(en^2)\mathcal{R}_n^2(\mathcal{F}(R_i))}\right).$$

This establishes one-sided subgaussian tail behavior. Now applying Proposition 11 and setting θ_i as suggested by the proposition we conclude that

$$\begin{aligned}
&\mathbb{E}_\epsilon\left[\max_{i\in\mathbb{N}} \sup_{f\in\mathcal{F}(R_i)} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon) - K_1 \mathcal{R}_n(\mathcal{F}(R_i)) \log^{3/2} n \left(1 + \sqrt{\log\left(\frac{\mathcal{R}_n(\mathcal{F}(R_i))}{\mathcal{R}_n(\mathcal{F}(1))}\right) + \log(\log(R_i))}\right)\right] \\
&\leq K_2 \Gamma \mathcal{R}_n(\mathcal{F}(1)) \log^{3/2} n.
\end{aligned}$$

This concludes the proof by appealing to Eq. (6.9). $\square$

Proof of Achievability for Example 8.

Lemma 11. The following bound is achievable in the setting of Example 8:

$$\mathcal{B}(f) = D\sqrt{n}\Big(8\|f\|\Big(1 + \sqrt{\log(2\|f\|) + \log\log(2\|f\|)}\Big) + 12\Big).$$

This proof specializes the proof of Corollary 8 to the regime where Lemma 8 applies.

Recall our parameterization of $\mathcal{F}$: $\mathcal{F}(R) = \{f \in \mathcal{F} : \|f\| \leq R\}$. It was shown in Rakhlin et al. (2012) that $\mathcal{C}_n(\mathcal{F}(R)) := 2RD\sqrt{n}$ is an upper bound for $\mathcal{R}_n(\mathcal{F}(R))$. We consider the rate

$$\mathcal{B}(f) = 2\mathcal{C}_n(\mathcal{F}(2R(f)))\left(1 + \sqrt{\log\left(\frac{\mathcal{C}_n(\mathcal{F}(2R(f)))}{\mathcal{C}_n(\mathcal{F}(1))}\right) + \log\log_2(2R(f))}\right).$$

We begin by applying Lemma 7, Eq. (6.3), yielding

$$\begin{aligned}
&\mathcal{V}_n^{\text{ol}}(\mathcal{F}, \mathcal{B}) \\
&\leq \sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \sup_{f} 2\sum_{t=1}^{n} \epsilon_t \langle f, \mathbf{y}_t(\epsilon)\rangle - 2\mathcal{C}_n(\mathcal{F}(2R(f)))\left(1 + \sqrt{\log\left(\frac{\mathcal{C}_n(\mathcal{F}(2R(f)))}{\mathcal{C}_n(\mathcal{F}(1))}\right) + \log\log_2(2R(f))}\right).
\end{aligned}$$

We now discretize the range of R via $R_i = 2^i$. By analogy with the proof of Corollary 8 we get the upper bound,

$$\begin{aligned}
&\sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \sup_{i \in \mathbb{N}} \left[\sup_{f \in \mathcal{F}(R_i)} 2\sum_{t=1}^{n} \epsilon_t \langle f, \mathbf{y}_t(\epsilon)\rangle - 2\mathcal{C}_n(\mathcal{F}(R_i))\left(1 + \sqrt{\log\left(\frac{\mathcal{C}_n(\mathcal{F}(R_i))}{\mathcal{C}_n(\mathcal{F}(1))}\right) + \log\log_2(R_i)}\right)\right] \\
&= \sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \sup_{i \in \mathbb{N}} \left[2R_i \left\|\sum_{t=1}^{n} \epsilon_t \mathbf{y}_t(\epsilon)\right\|_{\star} - 4D\sqrt{n}R_i\sqrt{\log(R_i) + \log(i)}\right].
\end{aligned}$$

Fix a $\mathcal{Y}$-valued tree $\mathbf{y}$ and define a set of random variables $X_i = 2R_i \|\sum_{t=1}^{n} \epsilon_t \mathbf{y}_t(\epsilon)\|_{\star}$. Let $B_i = 2D\sqrt{n}R_i$. Lemma 8 shows that

$$\mathbb{P}(X_i - B_i \geq \tau) \leq 2\exp\left(-\frac{\tau^2}{8D^2R_i^2 n}\right).$$

So we have $\sigma_i = 2DR_i\sqrt{n}$, and it will be sufficient to set $\bar{\sigma} = 2D\sqrt{n}$. Since our tail bound is purely sub-gaussian, we apply Proposition 11 with $\theta_i = \frac{\sigma_i}{B_i}\sqrt{2\log(\sigma_i/\bar{\sigma}) + 4\log(i)} + 1$, yielding the following bound:

$$\sup_{\mathbf{y}} \mathbb{E}_{\epsilon} \sup_{i \in \mathbb{N}} \left[2R_i \left\|\sum_{t=1}^{n} \epsilon_t \mathbf{y}_t(\epsilon)\right\|_{\star} - 4D\sqrt{n}R_i\sqrt{\log(R_i) + \log(i)}\right] \leq 12D\sqrt{n}.$$

□

Proof of Achievability for Example 11. Unfortunately, the general symmetrization proof in Lemma 7 does not suffice for this problem. In what follows we use a more specialized symmetrization technique to prove the lemma.

Lemma 12. For any countable class of experts, when we consider $\mathcal{F}$ to be the class of all distributions over the set of experts, the following adaptive bound is achievable:

$$\mathcal{B}(f;y_{1:n}) = \sqrt{50\left(\mathrm{KL}(f|\pi)+\log(n)\right)\sum_{t=1}^{n}\langle f, y_t\rangle} + 50\left(\mathrm{KL}(f|\pi)+\log(n)\right)+1.$$

To show that the rate is achievable we need to show that $\mathcal{V}_n^{\mathrm{ol}} \le 0$. Since each $\hat{y}_t$ is a distribution over experts and we are in the linear setting, we do not need to randomize in the definition of the minimax value. Let us use the shorthand

$$C(f) = \mathrm{KL}(f|\pi)+\log(n),$$

and take constants K_1, K_2 to be determined later. Define

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F},\mathcal{B})$$
$$= \left\langle\!\!\left\langle \inf_{\hat{y}_t\in\Delta}\sup_{y_t\in\mathcal{Y}} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n}\langle\hat{y}_t, y_t\rangle - \inf_{f\in\Delta}\left\{\sum_{t=1}^{n}\langle f, y_t\rangle + \sqrt{KC(f)\sum_{t=1}^{n}\mathop{\mathbb{E}}_{i\sim f}\langle e_i, y_t\rangle^2} + \sqrt{K'}C(f)\right\}\right].$$

Using repeated minimax swap, this expression is equal to

$$\left\langle\!\!\left\langle \sup_{p_t\in\Delta(\mathcal{Y})}\inf_{\hat{y}_t\in\Delta} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n}\langle\hat{y}_t, y_t\rangle - \inf_{f\in\Delta}\left\{\sum_{t=1}^{n}\langle f, y_t\rangle + \sqrt{KC(f)\sum_{t=1}^{n}\mathop{\mathbb{E}}_{i\sim f}\langle e_i, y_t\rangle^2} + \sqrt{K'}C(f)\right\}\right]$$
$$= \left\langle\!\!\left\langle \sup_{p_t\in\Delta(\mathcal{Y})}\mathop{\mathbb{E}}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n}\inf_{\hat{y}_t\in\Delta}\mathbb{E}_{y_t\sim p_t}\left[\langle\hat{y}_t, y_t\rangle\right]\right.$$
$$\left. - \inf_{f\in\Delta}\left\{\sum_{t=1}^{n}\langle f, y_t\rangle + \sqrt{KC(f)\sum_{t=1}^{n}\mathop{\mathbb{E}}_{i\sim f}\langle e_i, y_t\rangle^2} + \sqrt{K'}C(f)\right\}\right].$$

By sub-additivity of square-root we pass to an upper bound,

$$\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \inf_{\hat{y}_t \in \Delta} \mathbb{E}_{y_t \sim p_t} \left[\langle \hat{y}_t, y_t \rangle \right] - \mathbb{E}_{e_i \sim f} \left[\langle e_i, y_t \rangle \right] \right.$$
$$\left. - \sqrt{C(f) \left(K \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y_t \rangle^2 \right] + K' C(f) \right)} \right].$$

We now split the square root according to the formula $\sqrt{ab} = \inf_{\alpha > 0} \{a/2\alpha + \alpha b/2\}$ and note the range of the optimal value:

$$\frac{1}{\sqrt{n}} \leq \alpha^* = \sqrt{\frac{C(f)}{\left(K \sum_{t=1}^{n} \mathbb{E}_{i \sim f} \left[\langle e_i, y_t \rangle^2 \right] + K' C(f)\right)}} \leq \frac{1}{\sqrt{K'}}. \tag{6.10}$$

Let us discretize the interval by setting $\alpha_i = \frac{1}{\sqrt{K'}} 2^{-(i-1)}$ for $i = 1, \ldots, N$ and note that we only need to take $N = O(\log(n))$ elements. Write $I = \{\alpha_1, \ldots, \alpha_N\}$. Observe that

$$\sqrt{ab} = \inf_{\alpha > 0} \{a/2\alpha + \alpha b/2\} \geq \min_{\alpha \in I} \{a/4\alpha + \alpha b/2\}.$$

For the rest of the proof, the maximum over α is taken within the set I. We have

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}, \mathcal{B}) \leq \left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \Delta, \alpha} \sum_{t=1}^{n} \inf_{\hat{y}_t \in \Delta(\mathcal{F})} \mathbb{E}_{y_t} \left[\langle \hat{y}_t, y_t \rangle \right] - \mathbb{E}_{e_i \sim f} \left[\langle e_i, y_t \rangle \right] \right.$$
$$\left. - \frac{\alpha}{2} \left(K \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y_t \rangle^2 \right] + K' C(f) \right) - \frac{C(f)}{4\alpha} \right]. \tag{6.11}$$

Dropping some negative terms, we upper bound the last expression by

$$\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}, \alpha} \sum_{t=1}^{n} \langle f, \mathbb{E}\left[y'_t\right] - y_t \rangle - \frac{K\alpha}{2} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y_t \rangle^2 \right] - \frac{C(f)}{4\alpha} \right].$$

Adding and subtracting $\frac{\alpha}{4} \sum_{t=1}^{n} \mathbb{E}_{y'_t} \left[\mathbb{E}_{i \sim f} \left[\langle e_i, y'_t \rangle^2 \right] \right]$, the expression is at most

$$\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{f \in \mathcal{F}, \alpha} \sum_{t=1}^{n} \langle f, \mathbb{E}\left[y'_t\right] - y_t \rangle - \frac{K\alpha}{4} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y_t \rangle^2 \right] - \frac{K\alpha}{4} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{y'_t} \left[\mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y'_t \rangle^2 \right] \right] \right.$$
$$\left. + \frac{K\alpha}{4} \left(\sum_{t=1}^{n} \mathbb{E}_{y'_t} \left[\mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y'_t \rangle^2 \right] \right] - \mathop{\mathbb{E}}_{i \sim f} \left[\langle e_i, y_t \rangle^2 \right] \right) - \frac{C(f)}{4\alpha} \right].$$

Using Jensen's inequality to pull out expectations, we obtain an upper bound,

$$\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{y_t, y'_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \Bigg[\sup_{f \in \mathcal{F}, \alpha} \sum_{t=1}^{n} \langle f, y'_t - y_t \rangle - \frac{K\alpha}{4} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y_t \rangle^2\right] - \frac{K\alpha}{4} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y'_t \rangle^2\right] + \frac{K\alpha}{4} \left(\sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y'_t \rangle^2\right] - \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y_t \rangle^2\right] \right) - \frac{C(f)}{4\alpha} \Bigg].$$

Next, we introduce Rademacher random variables:

$$\begin{aligned} &\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{y_t, y'_t \sim p_t} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \Bigg[\sup_{f \in \mathcal{F}, \alpha} \sum_{t=1}^{n} \epsilon_t \bigg(\langle f, y'_t - y_t \rangle + \frac{K\alpha}{4} \Big(\mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y'_t \rangle^2\right] - \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y_t \rangle^2\right] \Big) \bigg) \\ &\qquad - \frac{K\alpha}{4} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y_t \rangle^2\right] - \frac{K\alpha}{4} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y'_t \rangle^2\right] - \frac{C(f)}{4\alpha} \Bigg] \\ &\leq \left\langle\!\!\left\langle \sup_{y_t} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \Bigg[\sup_{f \in \mathcal{F}, \alpha} \sum_{t=1}^{n} \epsilon_t \bigg(2\langle f, y_t \rangle + \frac{K\alpha}{2} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y_t \rangle^2\right] \bigg) - \frac{K\alpha}{2} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, y_t \rangle^2\right] - \frac{C(f)}{4\alpha} \Bigg]. \end{aligned}$$

Moving to the tree notation, we have

$$\begin{aligned} \sup_{\mathbf{y}} \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}, \alpha} \Bigg[&\sum_{t=1}^{n} \epsilon_t \bigg(2\langle f, \mathbf{y}_t(\epsilon) \rangle + \frac{K\alpha}{2} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, \mathbf{y}_t(\epsilon) \rangle^2\right] \bigg) \\ &- \frac{K\alpha}{2} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i \sim f}\left[\langle e_i, \mathbf{y}_t(\epsilon) \rangle^2\right] - \frac{\mathrm{KL}(f|\pi)}{4\alpha} - \frac{\log n}{4\alpha} \Bigg]. \end{aligned}$$

Let the tree $\mathbf{y}$ be fixed. Note that the convex conjugate of $\frac{1}{\alpha}\mathrm{KL}(f\|\pi)$ is given by $\Psi^*(X) := \frac{1}{\alpha} \log\left(\mathbb{E}_{i \sim \pi} \exp\left(\alpha \langle e_i, X \rangle\right)\right)$, so we can express the last quantity as

$$\mathop{\mathbb{E}}_{\epsilon} \max_{\alpha} \Bigg\{ \frac{1}{4\alpha} \log\Bigg(\mathop{\mathbb{E}}_{i \sim \pi} \exp\left(\sum_{t=1}^{n} \epsilon_t \big(8\alpha \langle e_i, \mathbf{y}_t(\epsilon) \rangle + 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon) \rangle^2 \big) - 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon) \rangle^2 \right) \Bigg) - \frac{\log n}{4\alpha} \Bigg\}.$$

Define a random variable indexed by α:

$$X_\alpha = \tfrac{1}{4\alpha} \log\left(\mathbb{E}_{i \sim \pi} \left[\exp\left(\textstyle\sum_{t=1}^{n} \epsilon_t \big(8\alpha \langle e_i, \mathbf{y}_t(\epsilon) \rangle + 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon) \rangle^2 \big) - 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon) \rangle^2 \right) \right] \right).$$

Our goal is to bound $\mathbb{E}\left[\max_\alpha \{X_\alpha - \log n / 4\alpha\}\right]$. Observe that the following chain of inequalities holds:

$$
\begin{aligned}
&\mathbb{P}(X_\alpha > t) \\
&\leq \inf_\lambda \mathbb{E}\left[e^{\lambda X_\alpha - \lambda t}\right] \\
&= \inf_\lambda \left\{ \mathop{\mathbb{E}}_{\epsilon} \left(\mathop{\mathbb{E}}_{i\sim\pi} \exp\left(\sum_{t=1}^n \epsilon_t \big(8\alpha\langle e_i, \mathbf{y}_t(\epsilon)\rangle + 2K\alpha^2\langle e_i, \mathbf{y}_t(\epsilon)\rangle^2\big) - 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon)\rangle^2 \right)\right)^{\frac{\lambda}{4\alpha}} e^{-\lambda t} \right\} \\
&\leq \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{i\sim\pi} \exp\left(\sum_{t=1}^n \epsilon_t \big(8\alpha\langle e_i, \mathbf{y}_t(\epsilon)\rangle + 2K\alpha^2\langle e_i, \mathbf{y}_t(\epsilon)\rangle^2\big) - 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon)\rangle^2 \right) e^{-4\alpha t} \\
&\leq \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{i\sim\pi} \exp\left(\sum_{t=1}^n \big(8\alpha\langle e_i, \mathbf{y}_t(\epsilon)\rangle + 2K\alpha^2\langle e_i, \mathbf{y}_t(\epsilon)\rangle^2\big)^2 - 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon)\rangle^2 \right) e^{-4\alpha t} \\
&\leq \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{i\sim\pi} \exp\left(\sum_{t=1}^n 4\alpha^2(4+K\alpha)^2\langle e_i, \mathbf{y}_t(\epsilon)\rangle^2 - 2K\alpha^2 \langle e_i, \mathbf{y}_t(\epsilon)\rangle^2 \right) e^{-4\alpha t}.
\end{aligned}
$$

The above term is upper bounded by $\exp(-4\alpha t)$ as soon as $4\alpha^2(4+K\alpha)^2 \leq 2K\alpha^2$, which happens when

$$0 < \alpha \leq (\sqrt{K/2} - 4)/K. \tag{6.12}$$

In view of (6.10), we know that $\alpha \leq \frac{1}{\sqrt{K'}}$. Thus, to ensure (6.12), it is sufficient to take $K = 50$ and $K' = 50^2$. Other choices lead to a different balance of constants. We thus have

$$\mathbb{P}(X_\alpha > t) \leq \exp(-4\alpha t).$$

Now that we have the tail bound, we appeal to Proposition 11. Setting $s_i = 4\alpha_i$ and $B_i = 1/4\alpha_i$, we obtain that

$$\mathbb{E}\left[\max_{i=1,\dots,N}\left\{X_{\alpha_i} - \frac{\log(n)}{4\alpha}\right\}\right] \leq 10.$$

□

6.6 Chapter Notes

This chapter is based on Foster et al. (2015).

Part III

New Guarantees for Adaptive Learning

CHAPTER 7

OVERVIEW OF PART III

In Part III of this thesis we apply the tools developed in Part II to four concrete settings of practical importance: online supervised learning, online convex optimization, statistical learning, and contextual bandits. For each setting we introduce a new type of adaptive learning guarantee, then use the equivalence framework to both develop efficient algorithms and characterize fundamental limits for the new guarantee.

- **Online supervised learning: Adaptivity to feature distribution.** In Chapter 8 we introduce a family of algorithms that adapt to the feature distribution in online learning with the property that their performance is *sequence optimal*, meaning no algorithm can obtain better statistical performance on *any sequence*. The achievability of this type of rate is characterized through a new connection to the theory of UMD Banach spaces.
- **Online convex optimization: Adaptivity to model.** In Chapter 9 we introduce the first universal family of *parameter-free* algorithms online and stochastic optimization. These algorithms learn the best learning rate or regularization parameter for the data, alleviating the need for parameter tuning.
- **Logistic regression: Adaptivity to misspecification.** In Chapter 10 we develop new

Setting	Adaptivity	Achievability	Algorithm
Online Learning	Feature distribution	Theorem 12	Theorem 11
Online Optimization	Model	Lemma 16/Theorem 22	Theorem 22
Statistical Learning	Misspecification	Theorem 35	Theorem 32
Contextual Bandits	Label distribution	Theorem 41	Theorem 42/43

Table 7.1: Summary of new adaptive learning algorithms and limits.

theory and algorithms for logistic regression in the presence of model misspecification. We design a new efficient *improper* learning algorithm for logistic regression that exhibits a *doubly-exponential* improvement in dependence certain parameters. This provides a positive resolution to a variant of the COLT 2012 open problem of McMahan and Streeter (2012), and also leads to the resolution of two open questions regarding adaptivity to margin in bandits and boosting. We use the minimax achievability framework to characterize the extent to which these improvements extend to general model classes.

- **Contextual bandits: Adaptivity to margin.** Chapter 11 introduces margin theory for contextual bandit learning. The new theory serves as a complete contextual bandit analogue of the classical margin theory in statistical learning, and permits the development of algorithms for sequential decision making that are more sample-efficient and computationally efficient when data is nice.

These results are summarized in Table 7.1.

CHAPTER 8

ONLINE SUPERVISED LEARNING

In this chapter we develop a new family of algorithms for the online learning setting with regret against any data sequence bounded by the *empirical Rademacher complexity* of the sequence. In the agnostic statistical learning setting, empirical Rademacher complexity (alongside the empirical covering numbers) is a fundamental quantity that adapts to the complexity of a benchmark function class $\mathcal{F}$ projected onto the observed dataset. We characterize when adaptivity based on empirical Rademacher complexity can be achieved in the more challenging online setting, and we derive efficient algorithms to achieve this.

Compared to the classical statistical setting, adaptivity to empirical Rademacher complexity is not always possible in the online setting. Achievability depends on refined properties of the benchmark class $\mathcal{F}$, and standard algorithms such as empirical risk minimization do not suffice to achieve this adaptivity. To characterize when this type of adaptive regret bound is achievable, we establish a connection to the theory of *decoupling inequalities* for martingales in Banach spaces. When the hypothesis class is a set of linear functions bounded in some norm, the empirical Rademacher complexity regret bound is achievable if and only if the norm satisfies certain decoupling inequalities (specifically, UMD inequalities) for martingales. In an instance of the equivalence framework of Part II, Donald Burkholder's celebrated *geometric characterization* of decoupling inequalities and UMD spaces (Burkholder, 1984) states that such an inequality holds if and only if there exists a *Burkholder function* satisfying a strengthening of the restricted concavity property called *zig-zag concavity*. Our online learning algorithms are efficient in terms of queries to this function.

We realize our general theory by giving new efficient and adaptive algorithms for linear classes including ℓ_p norms, group norms, and reproducing kernel Hilbert spaces, as well as adaptive regret guarantees for general classes based on empirical covering numbers. The empirical Rademacher complexity regret bound implies—when used in the i.i.d. setting—a *data-dependent* complexity bound for excess risk after online-to-batch conversion.

8.1 Background

We focus on the online supervised learning task (Protocol 2) for the special case of real-valued predictions. To recap, the learner receives data $(x_1, y_1), \ldots, (x_n, y_n)$ in a stream. At time t they receive an instance x_t and must predict y_t given the instance and the previous observations $(x_1, y_1,) \ldots, (x_{t-1}, y_{t-1})$. The learner's prediction, denoted $\widehat{y}_t$, is evaluated against y_t according to a loss function $\ell(\widehat{y}_t, y_t)$; for classification this is typically a convex surrogate for the zero-one loss $\ell_{01}(\widehat{y}, y) = \mathbb{1}\{\widehat{y} \neq y\}$ such as the hinge loss $\ell_{\text{hinge}}(\widehat{y}, y) = \max\{0, 1 - \widehat{y} \cdot y\}$. The learner's overall performance is measured in terms of their *regret* against a benchmark function class $\mathcal{F}$:

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t). \tag{8.1}$$

In the *statistical setting*, each pair (x_t, y_t) is drawn i.i.d. from some joint distribution $\mathcal{D}$. In this case, a bound on (8.1) is appealing because it immediately translates to an excess loss bound for the batch statistical learning setting after online-to-batch conversion. At the other extreme is the *fully adversarial* setting, where no generating assumptions on the data are made. This chapter develops methods that enjoy optimal guarantees in both worlds.

Our goal is to come up with prediction strategies that adapt to the "difficulty" of the sequence. In the statistical setting, optimal excess risk behavior has long been understood through empirical process theory and, in particular, Rademacher averages (Bartlett and Mendelson, 2003). Empirical Rademacher averages were shown to be an attractive data-dependent measure of complexity that can be used for model selection and for estimating the excess risk of empirical minimizers. The question considered in this chapter is whether there exist prediction strategies such that empirical Rademacher averages control the per-sequence regret (8.1). It turns out that the empirical Rademacher average is the best sequence-based measure of complexity one can hope for.

Let us formally define the *empirical Rademacher complexity* of the class $\mathcal{F}$:

$$\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}) = \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t), \tag{8.2}$$

where the Rademacher sequence $\epsilon \in \{\pm 1\}^n$ is drawn uniformly at random and $x_{1:n} = (x_1, \ldots, x_n)$. The questions studied in this chapter are:

- *When does there exist a strategy* $(\widehat{y}_t)$ *such that*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \mathbf{D}(\mathcal{F}, n) \cdot \widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}) \tag{8.3}$$

  for every sequence $x_{1:n}, y_{1:n}$*?*

- *What is the best constant* $\mathbf{D}(\mathcal{F}, n)$*?*

- *When can the strategy* $(\widehat{y}_t)_{t \geq 1}$ *be efficiently computed?*

We provide a characterization of when the bound (8.3) is achievable, and, furthermore, develop efficient algorithms based on a new set of techniques. This is achieved using a specialized version of the equivalence developed in Part II. Interestingly, the Burkholder functions that arise from the equivalence in this section satisfy a stronger form of restricted concavity called "zig-zag concavity" (see Figure 8.1). The main message of this chapter is that this special function can be used for algorithmic purposes and to answer the above questions.

We begin our analysis by showing that the empirical Rademacher complexity $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n})$ enjoys a rather strong form of instance optimality that we term "sequence optimality". This result is a simple consequence of the equivalence of martingale inequalities and adaptive prediction guarantees.

Lemma 13 (Sequence Optimality)**.** Let ℓ be the absolute, hinge, or linear loss and let $\mathcal{F}$ be any class of functions with value bounded by 1. Let $\mathcal{B}(x_{1:n})$ be a data-dependent regret bound for which there exists a strategy $(\widehat{y}_t)$ guaranteeing

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \mathcal{B}(x_{1:n}) \qquad \forall x_{1:n}, y_{1:n}. \tag{8.4}$$

Then

$$\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}) \leq \mathcal{B}(x_{1:n}) \quad \forall x_{1:n}.$$

The same result holds for the zero-one loss if we restrict $\mathcal{F}$ and $(\widehat{y}_t)$ to have range $\{\pm 1\}$.

Lemma 13 reveals that no data-dependent regret bound can improve upon $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n})$ beyond the factor $\mathbf{D}(\mathcal{F}, n)$. As we will soon show, the question of identifying $\mathbf{D}(\mathcal{F}, n)$ is an extremely rich one. When one restricts to linear function classes, this question is deeply tied to theory of Banach spaces with the *unconditional martingale difference* (UMD) property.

For the majority of this chapter we assume that $\mathcal{F}$ is a class of linear functions indexed by a unit ball; Section 8.6 considers the general case. For the linear case, we assume that x_ts lie in the unit ball of a separable Banach space $(\mathfrak{B}, \|\cdot\|)$ and

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid w \in \mathfrak{B}^\star, \|w\|_\star \leq 1\},$$

with $\|\cdot\|_\star$ being the dual norm and $\mathfrak{B}^\star$ the dual space. In this case,

$$\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}) = \mathop{\mathbb{E}}_{\epsilon} \sup_{\|w\|_\star \leq 1} \sum_{t=1}^{n} \epsilon_t \langle w, x_t \rangle = \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|.$$

Consider the euclidean setting, where $\mathcal{F}$ is the unit ℓ_2 ball. It is known that gradient descent with an adaptive step size yields a regret bound of order $\sqrt{\sum_{t=1}^{n} \|x_t\|^2}$ for any sequence.[1] Khintchine's inequality gives a further upper bound of order $\mathbb{E}_\epsilon \|\sum_{t=1}^{n} \epsilon_t x_t\|$. Hence, adaptive gradient descent answers the questions posed earlier for the specific case of linear functions indexed by Euclidean ball. This is one of two cases known to us where the bound of $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n})$ was available without using the techniques developed in this chapter.[2]

8.2 Preliminaries

Let $(\mathfrak{B}, \|\cdot\|)$ be a separable Banach space and $(\mathfrak{B}^\star, \|\cdot\|_\star)$ denote its dual. This chapter focuses on the real-valued online supervised learning setting described in Section 2.3. Input instances belong to some subset $\mathcal{X} \subseteq \mathfrak{B}$ and predictions $\widehat{y}_t$ are real valued ($\widehat{\mathcal{Y}} = \mathbb{R}$). The outcomes ($y_t$s) are selected from some abstract label space $\mathcal{Y}$. Throughout the chapter we assume that the loss $\ell(\hat{y}, y)$ is convex and 1-Lipschitz in its first argument. We also assume that there exists some bounded domain $[-B, B]$ such that for all $y \in \mathcal{Y}$, $\exists \hat{y} \in [-B, B]$ such that the derivative with respect to the first argument $\ell'(\hat{y}, y) = 0$ (that is, minimum is achievable in the compact set). We call such a loss function *well-behaved*. We remark that this bound B never explicitly appears in our results, and its only purpose is to enable application of the minimax theorem (Section 2.6).

Additional Notation For $p \in (1, \infty)$, let $p' = p/(p-1)$ denote its conjugate, and $p^\star = \max\{p, p'\}$. For a matrix $X \in \mathbb{R}^{d \times d}$, let $X_{i,\cdot}$ denote the ith row and $X_{\cdot j}$ denote the jth column. We define its (p, q) group norm as $\|X\|_{p,q} = (\sum_{i \in [d]} \|X_{i,\cdot}\|_q^p)^{1/p} = \|(\|X_{i,\cdot}\|_q)_{i \in [d]}\|_p$.

[1] See discussion in Chapter 5.
[2] The other example is the ℓ_∞ ball, attained by diagonal AdaGrad (Duchi et al., 2011).

For a set $\mathcal{A} \subseteq \mathbb{R}^d$, assumed to be symmetric, the atomic norm with respect to $\mathcal{A}$ is given by $\|x\|_{\mathcal{A}} = \min\{\alpha \mid x \in \alpha \cdot \mathrm{conv}(\mathcal{A})\}$.

We use the convention that both $\epsilon \in \{\pm 1\}^n$ and $\sigma \in \{\pm 1\}^n$ denote Rademacher sequences.

8.3 Burkholder Method and Zig-Zag Concavity

Let us propose a simple schema for designing algorithms to achieve (8.3). It will turn out that considering this scheme naturally leads to us to decoupling inequalities for Banach space-valued martingales via Burkholder's method. We begin by observing that by convexity of the loss function,

$$\ell(\widehat{y}_t, y_t) - \ell(\langle w, x_t \rangle, y_t) \leq \ell'(\widehat{y}_t, y_t) \cdot (\widehat{y}_t - \langle w, x_t \rangle) \tag{8.5}$$

and hence, denoting the derivative by $\ell'_t = \ell'(\widehat{y}_t, y_t)$,

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{\|w\|_\star \leq 1} \sum_{t=1}^{n} \ell(\langle w, x_t \rangle, y_t) \leq \sum_{t=1}^{n} \widehat{y}_t \cdot \ell'_t + \left\| \sum_{t=1}^{n} \ell'_t x_t \right\|. \tag{8.6}$$

Rather than directly aiming for the adaptive bound of empirical Rademacher averages in (8.3), we shall aim for $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}, \ell'_{1:n}) := \mathbb{E}_\epsilon \|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\|$, a quantity that is always tighter than $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}) = \mathbb{E}_\epsilon \|\sum_{t=1}^{n} \epsilon_t x_t\|$ because ℓ is 1-Lipschitz.

Consequently, to achieve the adaptive regret bound (8.3), it suffices to exhibit a strategy for which the quantity

$$\sum_{t=1}^{n} \widehat{y}_t \cdot \ell'_t + \left\| \sum_{t=1}^{n} \ell'_t x_t \right\| - \mathbf{D} \cdot \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \ell'_t x_t \right\|$$

is at most zero on every data sequence.

The challenge in analyzing this quantity is that the function $z \mapsto \|A + z\| - \mathbf{D}\|B + \epsilon z\|$ is neither convex nor concave. Virtually all potential functions used in online learning are convex and the absence of such a property makes it difficult to bound the growth under possible outcomes for the gradient ℓ'_t. Thankfully, the Burkholder method suggests an extremal function that enjoys more favorable analytical properties.

Proposition 12. Suppose there exists a function $\mathbf{U} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ satisfying

1. $\mathbf{U}(x, x') \geq \|x\| - \mathbf{D}\|x'\|$.
2. $\mathbf{U}$ is ***zig-zag concave***: $z \mapsto \mathbf{U}(x + z, x' + \epsilon z)$ is concave for all $x, x' \in \mathfrak{B}$ and $\epsilon \in \{\pm 1\}$.
3. $\mathbf{U}(0, 0) \leq 0$.

Then the simple gradient-based strategy

$$\widehat{y}_t = -\frac{d}{d\alpha} \mathop{\mathbb{E}}_{\epsilon_{1:t}} \mathbf{U}\left(\sum_{s=1}^{t-1} \ell'_s x_s + \alpha x_t, \sum_{s=1}^{t-1} \epsilon_s \ell'_s x_s + \epsilon_t \alpha x_t\right)\Bigg|_{\alpha=0} \tag{8.7}$$

achieves the empirical Rademacher complexity regret bound (8.3).

We remark that this strategy is horizon-independent whenever $\mathbf{U}$ does not depend on n (which is the case for examples we consider). Furthermore, one may avoid re-drawing the random signs and, hence, the computation time is simply the evaluation of the derivative of $\mathbf{U}$. As a consequence, the sufficient statistics for adapting to the empirical Rademacher complexity are simply the sequence $\sum \ell'_t x_t$ and its sign-flipped cousin $\sum \epsilon_t \ell'_t x_t$.

The full description this strategy, which we call the ZigZag algorithm is given in Section 8.5. We postpone this discussion for a moment in favor of connecting the zig-zag concave Burkholder functions to other properties of the Banach space via the equivalence.

8.4 Zig-Zag Functions, Regret, and UMD Spaces

What have we gained by reducing our problem to finding a $\mathbf{U}$ function? We will now show that $\mathbf{U}$ exists *if and only if* $(\mathfrak{B}, \|\cdot\|)$ is an *Unconditional Martingale Difference* (UMD) space. Informally, in a UMD space lengths of martingales are comparable to those of random walks with independent increments (see Definition 4). We call $\mathbf{U}$ a *Burkholder function* in reference to Donald Burkholder's central result characterizing UMD spaces in terms of the existence of these functions (Burkholder, 1984).

In Proposition 12 we assumed that the Burkholder function $\mathbf{U}$ satisfies $\mathbf{U}(x, x') \geq \|x\| - \mathbf{D}\|x'\|$. We will soon see that it is often easier to find an efficiently computable zig-zag concave function $\mathbf{U}_p$ that, as before, satisfies $\mathbf{U}_p(0,0) \leq 0$, but the first requirement in Proposition 12 is replaced with

$$\mathbf{U}_p(x, x') \geq \|x\|^p - \mathbf{D}_p^p \|x'\|^p$$

for some $p > 1$ (i.e. $p \neq 1$). However, the simple observation that for any number $a > 0$, $a = \frac{1}{p} \inf_{\eta > 0}\{\eta a^p + (p-1)\eta^{-1/(p-1)}\}$ will allow us to algorithmically use a $\mathbf{U}_p$ function for any p to obtain the desired regret bound $\widehat{\mathcal{R}}$ (this is described in detail in Section 8.5). This motivates our complete Burkholder function definition:

Definition 3. *A function* $\mathbf{U}_p^{\mathfrak{B}} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ *is* Zig-Zag *for* $(\|\cdot\|, p, \mathbf{D}_p)$ *if*

1. $\mathbf{U}_p^{\mathfrak{B}}(x, x') \geq \|x\|^p - \mathbf{D}_p^p \|x'\|^p$.
2. $\mathbf{U}_p^{\mathfrak{B}}$ *is* ***zig-zag concave****: The function* $z \mapsto \mathbf{U}_p^{\mathfrak{B}}(x + z, x' + \epsilon z)$ *is concave for all* $x, x' \in \mathfrak{B}$ *and* $\epsilon \in \{\pm 1\}$.
3. $\mathbf{U}_p^{\mathfrak{B}}(0,0) \leq 0$.[3]

[3]This condition is without loss of generality.

For concreteness, here is a simple example for the scalar case: The function

$$\mathbf{U}_2^{\mathbb{R}}(x, x') = |x|^2 - |x'|^2$$

is Zig-Zag for $(|\cdot|, 2, 1)$. The reader can easily verify that this function is zig-zag concave by observing that $\mathbf{U}_2^{\mathbb{R}}(x + z, x' \pm z)$ is in fact linear in z. Perhaps the most famous $\mathbf{U}$ function is Burkholder's construction for general powers in the scalar case: For $p \in (1, \infty)$ the function

$$\mathbf{U}_p^{\mathbb{R}}(x, x') = \alpha_p(|x| - \beta_p|x'|)(|x| + |x'|)^{p-1},$$

is a $(|\cdot|, p, \beta_p)$ Burkholder function upper bounding $|x|^p - \beta_p^p|x'|^p$ for appropriate α_p, β_p.

8.4.1 When Does a Zig-Zag Concave Burkholder Function Exist?

It turns out that the most common Banach spaces used in machine learning settings — such as ℓ_p spaces, group norms, Schatten-p classes, and operator norms — all happen to be UMD spaces, and that each UMD space comes with its own $\mathbf{U}$ function. This leaves us with the exciting prospect of using their corresponding $\mathbf{U}$ functions to develop new adaptive online learning algorithms with improved data-dependent regret bounds. Without further ado, let us define a UMD Banach space:

Definition 4. *A Banach space $(\mathfrak{B}, \|\cdot\|)$ is called UMD$_p$ for some $1 < p < \infty$, if there is a constant $\mathbf{C}_p$ such that for any finite $\mathfrak{B}$-valued martingale difference sequence $(X_t)_{t=1}^n$ in $L_p(\mathfrak{B})$ and any fixed choice of signs $(\epsilon_t)_{t=1}^n$ (where each $\epsilon_t \in \{\pm 1\}$),*

$$\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t X_t\right\|^p \leq \mathbf{C}_p^p\, \mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|^p. \tag{8.8}$$

The space $(\mathfrak{B}, \|\cdot\|)$ is called UMD_1 *if there is a constant* $\mathbf{C}_1$ *such that*

$$\mathbb{E}\sup_{\tau \leq n}\left\|\sum_{t=1}^{\tau} \epsilon_t X_t\right\| \leq \mathbf{C}_1\, \mathbb{E}\sup_{\tau \leq n}\left\|\sum_{t=1}^{\tau} X_t\right\|. \tag{8.9}$$

Burkholder (1984) proved the following geometric characterization of UMD spaces in terms of existence of appropriate zig-zag concave $\mathbf{U}$ functions.[4]

Theorem 7 (Hytönen et al. (2016), Theorem 4.5.6)**.** *For a Banach space $(\mathfrak{B}, \|\cdot\|)$, the following are equivalent:*

1. *$\mathfrak{B}$ is UMD$_p$ with constant $\mathbf{C}_p$.*
2. *There exists Burkholder function $\mathbf{U}_p^{\mathfrak{B}} : \mathfrak{B} \times \mathfrak{B} \mapsto \mathbb{R}$ for $(\|\cdot\|, p, \mathbf{C}_p)$.*

Theorem 7 is strengthened considerably by the following fact:

Theorem 8. *Let $p \in (1, \infty)$. If UMD_p holds with constant $\mathbf{C}_p$, then*

- *For all $q \in (1, \infty)$, UMD_q, holds with constant $\mathbf{C}_q \leq 100\left(\frac{q}{p} + \frac{q'}{p'}\right)\mathbf{C}_p$.*
- *UMD_1 holds with $\mathbf{C}_1 = O(\mathbf{C}_p)$.*

Furthermore, if UMD_1 holds with constant $\mathbf{C}_1$, then for all $p \in (1, \infty)$ there is some constant $\mathbf{C}'_p$ for which UMD_p holds.

With these properties of UMD spaces established, we proceed to state our main theorem on achieving the $\widehat{\mathcal{R}}$ regret bound in these spaces.

Theorem 9. *Let $(\mathfrak{B}, \|\cdot\|)$ satisfy UMD_p with constant $\mathbf{C}_p$ for any $p \in [1, \infty)$. Then there exists*

[4]Burkholder (1984) does not work with $\mathbf{U}$ functions directly but rather an equivalent property called ζ-convexity. The $\mathbf{U}$ function presentation first appeared in Burkholder (1986). See Hytönen et al. (2016) or Osekowski (2012) for a modern exposition.

some randomized strategy achieving the regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right]\leq O\left(\mathbf{C}_p\,\mathbb{E}\,\mathbb{E}_{\epsilon}\sup_{\tau\leq n}\left\|\sum_{t=1}^{\tau}\epsilon_t\ell'(\widehat{y}_t,y_t)x_t\right\|\right) \tag{8.10}$$

$$\leq O\left(\mathbf{C}_p\,\mathbb{E}\left(\log\left(\max_{t\in[n]}\|x_t\|n\right)\cdot\mathbb{E}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\ell'(\widehat{y}_t,y_t)x_t\right\|\right)\right) \tag{8.11}$$

$$\leq O\left(\mathbf{C}_p\,\mathbb{E}\left(\log\left(\max_{t\in[n]}\|x_t\|n\right)\cdot\mathbb{E}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t x_t\right\|\right)\right). \tag{8.12}$$

This shows that a bound on $\mathbf{C}_p$ for any p gives $\mathbf{D}(\mathcal{F},n)\leq\mathbf{C}_p$ in (8.3), up to an extra additive $\log n$ factor[5].

An interesting feature of this theorem is that there are multiple ways through which it can be proven. In Section 8.7 it is proven purely *non-constructively* by plugging the UMD inequality (8.9) into the minimax analysis framework developed in Foster et al. (2015). In Section 8.5 it is proven *constructively* by using the existence of the $\mathbf{U}$ function to exhibit a particular strategy for the learner.

Let us remark that the bound in (8.10) has the desirable property of adapting to scale, in that it does not require an a-priori upper bound on the data norms $\max_{t\in[n]}\|x_t\|$.

With Theorem 9 in mind, we finally state bounds on $\mathbf{C}_p$ for classes of interest.

Theorem 10. *The following* UMD *constants hold:*

[5]All of the $\log n$ factors incurred in this paper arise when passing from bounds of the form $\mathbb{E}\sup_{\tau\leq n}F_\tau$ to those of the form $\mathbb{E}\,F_n$ for some random process (F_t). This is notable technical issue with most martingale inequalities involving the $L_1(\mathfrak{B})$ norm, including for example Doob's well-known maximal inequality.

- $(\mathbb{R}, |\cdot|)$: $\mathbf{C}_p = p^\star - 1\ \forall p \in (1,\infty)$.
- $(\mathbb{R}^d, \|\cdot\|_p)$, $p \in (1,\infty)$: $\mathbf{C}_p = p^\star - 1$.
- $(\mathbb{R}^d, \|\cdot\|_1/\|\cdot\|_\infty)$: $\mathbf{C}_2 = O(\log d)$.
- $(\mathbb{R}^d, \|\cdot\|_{\mathcal{A}}/\|\cdot\|_{\mathcal{A}^\star})$: $\mathbf{C}_2 = O(\log|\mathcal{A}|)$.
- $(\mathbb{R}^{d\times d}, \|\cdot\|_{S_p})$, $p \in (1,\infty)$: $\mathbf{C}_p = O((p^\star)^2)$.
- $(\mathbb{R}^{d\times d}, \|\cdot\|_\sigma/\|\cdot\|_\Sigma)$: $\mathbf{C}_2 = O(\log^2 d)$.
- $(\mathbb{R}^{d\times d}, \|\cdot\|_{p,q})$, $p, q \in (1,\infty)$: $\mathbf{C}_p = O(p^\star q^\star)$.
- $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ *for Hilbert space* $\mathcal{H}$: $\mathbf{C}_2 = 1$.

8.4.2 Efficient Burkholder Functions

Burkholder's geometric characterization, Theorem 7, implies existence of a Burkholder function $\mathbf{U}_p^{\mathfrak{B}}$ whenever a space $(\mathfrak{B}, \|\cdot\|)$ has UMD constant $\mathbf{C}_p$. Unfortunately, the generic $\mathbf{U}$ function construction (see Hytönen et al. (2016), Theorem 4.5.6) is not *efficiently computable*; it is expressed in terms of a supremum over all martingale difference sequences. However, the construction of concrete $\mathbf{U}$ functions has been an active area of research in the three decades since Burkholder's original construction. This is because one can exhibit a $\mathbf{U}$ function to certify that a space is UMD for a specific constant $\mathbf{C}_p$, and discovering *sharp* UMD constants is of general interest to the analysis community (Osekowski, 2012).

Let us begin by stating Burkholder's optimal $\mathbf{U}$ function construction for the scalar setting. This function was originally obtained by solving a particular partial differential equation. This function is graphed in Figure 8.1.

Example 12 ($|\cdot|^p$, Hytönen et al. (2016), Theorem 4.5.7)**.** *For any* $p \in (1,\infty)$, *the function*

$$\mathbf{U}_p^{\mathbb{R}}(x,y) := \alpha_p(|x| - \beta_p|y|)(|x|+|y|)^{p-1} \tag{8.13}$$

is Zig-Zag for $(|\cdot|, p, \beta_p)$, *where* $\alpha_p = p\left(1 - \frac{1}{p^\star}\right)^{p-1}$, $\beta_p = p^\star - 1$. β_p *is the sharpest constant possible.*

Observe that all of the Burkholder function properties (Definition 3) are preserved under addition. This leads us to a construction for ℓ_p norms in the vector setting, which inherits the optimal constants from Burkholder's scalar construction.

Example 13 (ℓ_p norm)**.**

$$\mathbf{U}_p^{\ell_p}(x, y) := \sum_{i \in [d]} \mathbf{U}_p^{\mathbb{R}}(x_i, y_i) \tag{8.14}$$

is a Burkholder function for $(\|\cdot\|_p^p, p, \beta_p)$*, with* β_p *as in* Example 12*.* $\mathbf{U}_p^{\ell_p}$ *can be computed in time* $O(d)$*.*

Example 14 (Weighted ℓ_2 norm)**.** *Let* $\|x\|_A = \sqrt{\langle x, Ax \rangle}$ *for some PSD matrix* A*. Then*

$$\mathbf{U}_2^{\ell_{2,A}}(x, y) := U_2^{\ell_2}(A^{1/2}x, A^{1/2}y)$$

is a Burkholder function for $(\ell_{2,A}, 2, 1)$*.* $\mathbf{U}_2^{\ell_{2,A}}$ *can be computed in time* $O(d^2)$*.*

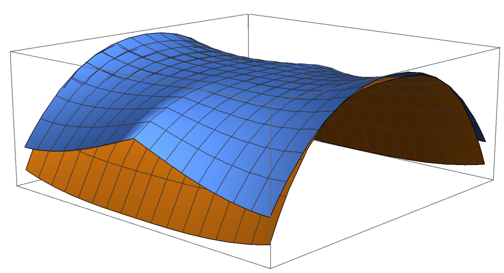


Figure 8.1: $\mathbf{U}_p^{\mathbb{R}}(x, x')$ (blue) and $|x|^p - \beta_p^p |x'|^p$ (orange) for $p = 3$.

Another useful construction extends Burkholder's scalar function to general Hilbert spaces. This is useful as it applies even to infinite dimensional spaces such as RKHS.

Example 15 (General Hilbert Space, Hytönen et al. (2016), Theorem 4.5.14)**.** *Let* $\mathcal{H}$ *be some Hilbert space whose norm will be denoted* $\|\cdot\|_{\mathcal{H}}$*.*

$$\mathbf{U}_p^{\mathcal{H}}(x, y) := \alpha_p(\|x\|_{\mathcal{H}} - \beta_p\|y\|_{\mathcal{H}})(\|x\|_{\mathcal{H}} + \|y\|_{\mathcal{H}})^{p-1} \tag{8.15}$$

is a Burkholder function for $(\|\cdot\|_{\mathcal{H}}, p, \beta_p)$ *for each* $p \in (1, \infty)$*, where* α_p *and* β_p*, and are as in* Example 12*. This function works for all Hilbert spaces, even those of infinite dimension. For* $p = 2$ *this function and its derivatives can be implemented efficiently using the Representer Theorem.*

We can lift the former construction to a construction for group norms in the same fashion as in our construction for ℓ_p norms.

Example 16 ($(p, 2)$ Group Norm)**.** *In this example we consider group norms over matrices in $\mathbb{R}^{d\times d}$. The function,*

$$\mathbf{U}_p^{(p,2)}(x, y) := \sum_{i\in[d]} \mathbf{U}_p^{\ell_2}(x, y),$$

where $\mathbf{U}^{\ell_2,p}$ is the general Hilbert space Burkholder function (8.15), is a Burkholder function for $(\|\cdot\|_{(p,2)}, p, \beta_p)$. $\mathbf{U}_p^{(p,2)}$ can be computed in time $O(d^2)$.

Group norms are used in multi-task learning. Furthermore, Example 16 works not just for $\mathbb{R}^{d\times d}$, but more generally for $\mathbb{R}^d \times \mathcal{H}$ for any Hilbert space $\mathcal{H}$. This makes it well-suited to multiple kernel learning tasks.

As we will show in the sequel, there are a number of algorithmic tricks we can use to adapt to the empirical Rademacher complexity even when we do not exactly have a zig-zag concave Burkholder function for the class of interest.

8.5 Algorithm and Applications

Recall that our goal is to design algorithms whose regret is bounded by $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}, \ell'_{1:n}) = \mathbb{E}_\epsilon\|\sum_{t=1}^n \epsilon_t \ell'_t x_t\|$. We now present an algorithm, ZIGZAG (Algorithm 5), which efficiently achieves a regret bound of this form whenever we have an efficient Burkholder function $\mathbf{U}_p^{\mathfrak{B}}$, even if $p \neq 1$.

Theorem 11. *Denote the prediction of Algorithm 5 as $\widehat{y}_t^{\epsilon_{1:t-1}}$ to make the dependence on the*

Algorithm 5 ZIGZAG

procedure ZIGZAG($\mathbf{U}_p, p, \eta$) ▷ $\mathbf{U}_p$ is Zig-Zag for $(\|\cdot\|, p, \beta)$. $\eta > 0$ is the learning rate.
  for time $t = 1, \ldots, n$ **do**
    Let $G_t(\alpha) = \mathbb{E}_{\sigma_t \in \{\pm 1\}} \frac{\eta}{p} \mathbf{U}_p\left(\sum_{s=1}^{t-1} \ell'_s x_s + \alpha x_t, \sum_{t=1}^{t-1} \epsilon_s \ell'_s x_s + \sigma_t \alpha x_t\right)$.
    Predict $\widehat{y}_t = -G'_t(0)$. ▷ More generally, use the supergradient.
    Draw independent Rademacher $\epsilon_t \in \{\pm 1\}$.
  end for
end procedure

sequence $(\epsilon_t)_{t \le n}$ explicit. Algorithm 5 enjoys the regret bound,

$$\mathbb{E}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \frac{1}{p}\left(\eta \beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p + \frac{1}{p'-1}\eta^{-(p'-1)}\right)\right] \le 0. \tag{8.16}$$

A few remarks are in order. A naive application of the Burkholder algorithm would yield a bound

$$\mathbb{E}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \le \frac{1}{p}\left(\eta \beta^p \mathbb{E}_{\epsilon} \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p + \frac{1}{p'-1}\eta^{-(p'-1)}\right), \tag{8.17}$$

which falls short of the goal of achieving $\widehat{\mathcal{R}}$ for the following reason. Observe that for any $p > 1$,

$$x^{1/p} = \frac{1}{p} \inf_{\eta > 0}\left(\eta x + \frac{1}{p'-1}\eta^{1-p'}\right) := \inf_{\eta > 0} \Psi_{\eta,p}(x). \tag{8.18}$$

Recall that $\eta > 0$ is a parameter of Algorithm 5. (8.18) combined with (8.17) suggest that if we chose the optimal η in hindsight, the regret of ZIGZAG would be bounded by $\sqrt[p]{\mathbb{E}_{\epsilon}\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\|^p}$. However, this bound is always worse than $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}, \ell'_{1:n})$ via Jensen's inequality, and is indeed sub-optimal for ℓ_p norms. Luckily, (8.16) reveals that for ZIGZAG, the Rademacher sequence $(\epsilon_t)_{t \le n}$ used by the algorithm and the Rademacher sequence appearing in the regret bound are one and the same, which allows us to adapt η to $\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\|$ for a particular playout of the sequence $(\epsilon_t)_{t \le n}$ to get the desired empirical Rademacher complexity bound. This tuning of η via doubling is stated in the next result.

Lemma 14. Define

$$\Phi(x_{t_1:t_2}, , \ell'_{t_1:t_2}, \epsilon_{t_1:t_2}) = \beta^p \sup_{t_1 \le a \le b \le t_2} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\|^p.$$

Consider the following strategy:

1. Choose $\eta_0 = (\beta \cdot p)^{-p}$ for $p \ge 2$ and $\eta_0 = 1$ for $p < 2$. Update with $\eta_i = 2^{-\frac{i}{p'-1}} \eta_0$.
2. In phase i, which consists of all $t \in \{s_i, \dots, s_{i+1} - 1\}$, play Algorithm 5, ZIGZAG, with learning rate η_i.
3. Take $s_1 = 1$, $s_{N+1} = n + 1$, and $s_{i+1} = \inf\{\tau \mid \eta_i \Phi(x_{s_i:\tau-1}, \ell'_{s_i:\tau-1}, \epsilon_{s_i:\tau-1}) > \eta_i^{-(p'-1)}\}$, where N is the index of the last phase (note that whether $t = s_{i+1}$ can be tested using only information available to the learner at time t).

This strategy achieves

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \tag{8.19}$$

$$\le O\left(\beta^2 \log^2 n \mathop{\mathbb{E}}_{\epsilon, \epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \ell'_t x_t \right\| + \min\left\{\log n + (p \cdot \beta)^{\frac{p}{p-1}}, \beta^p \log n\right\}\right). \tag{8.20}$$

Remark 1. *In the above bound, (x_t) and (ℓ'_t) may adapt to the sequence (ϵ_t) drawn by the algorithm (unless the adversary is oblivious), but may not adapt to (ϵ'_t).*

8.5.1 ℓ_p norms

We now specialize our generic algorithm to the important special case of ℓ_p norms. We use $\mathbb{E}$ (without subscript) to denote the expectation with respect to the learner's randomization.

Example 17. *Fix $p \in (1, \infty)$. Let $\widehat{y}_t$ be the strategy produced by* ZIGZAG *(Algorithm 5) using the Burkholder function $\mathbf{U}_p^{\ell_p}$ from Example 13 with the learning rate tuning strategy from Lemma 14.*

This strategy achieves

$$\mathbb{E}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right]\leq O\left(\mathbb{E}\,\mathbb{E}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t\ell_t'x_t\right\|_p\cdot(p^\star)^2\log^2 n+(p^\star)^2\log n\right). \quad (8.21)$$

This algorithm serves as a generalization of AdaGrad to all powers of p. If we take $p=2$, the result recovers the regret bound for full matrix AdaGrad (Duchi et al., 2011) up to logarithmic factors:

$$\mathbb{E}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right]\leq \tilde{O}\left(\mathbb{E}\sqrt{\sum_{t=1}^{n}\|x_t\|_2^2}\right). \quad (8.22)$$

We can also recover the regret bound for diagonal AdaGrad (Duchi et al., 2011) by taking $p=1+1/\log d$:

$$\mathbb{E}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right]\leq \tilde{O}\left(\mathbb{E}\sum_{i\in[d]}\|x_{1:n,i}\|_2\right). \quad (8.23)$$

Here $x_{1:n,i}$ denotes the ith row of the data matrix $(x_1,x_2,\dots,x_n)\in\mathbb{R}^{d\times n}$

There is also a direct construction of a $\mathbf{U}$ function for ℓ_1 due to Osekowski (2016), which is stated in Section 8.7 as Example 20. Using this function we will achieve (8.23), but without having to use the learning rate tuning strategy, and with only $O(\log d)$ factors in the regret bound instead of $O(\log^2 d)$.

8.6 Beyond Linear Function Classes: Necessary and Sufficient Conditions

The aim this chapter was to analyze conditions for the existence of adaptive methods that enjoy per-sequence empirical Rademacher complexity as the regret bound. In this quest, we introduced the UMD property as a necessary condition. In the present section, we consider arbitrary, possibly nonlinear function classes $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and show that a closely related "probabilistic" UMD property offers both a necessary *and* sufficient condition.

For this section we restrict ourselves to absolute loss $\ell_{\mathrm{abs}}(\hat{y}, y) = |\hat{y} - y|$ and assume that $\mathcal{Y} = [-1,1]$.

Theorem 12. *Let $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$ be any class of predictors. The following statements are equivalent:*

1. *There exists a learning algorithm and constant B such that the following regret bound holds against any adversary:*

$$\sum_{t=1}^{n} \ell_{\mathrm{abs}}(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell_{\mathrm{abs}}(f(x_t), y_t) \leq B \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) + b.$$

2. *For any $\mathcal{X}$ valued tree $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where each $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$, there exists constant C such that*

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1}))\right] \leq C \mathop{\mathbb{E}}_{\epsilon, \epsilon'}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))\right] + c, \tag{8.24}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ and $\epsilon' = (\epsilon'_1, \ldots, \epsilon'_n)$ are independent Rademacher random variables.

Moreover, $B = \Theta(C)$ and $b = \Theta(c)$. More generally, condition 2 implies condition 1 for any loss ℓ that is 1-Lipschitz and well-behaved as in Section 8.2, for any choice of $\mathcal{Y}$.

8.6.1 Function Classes with the Generalized UMD Property

We now give examples of function classes that satisfy the generalized UMD inequality (8.24).

Example 18 (Kernel Classes)**.** *Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space with kernel K such that $\sup_{x \in \mathcal{X}} \sqrt{K(x,x)} \leq B$, and let $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$. Then there are constants K_1, K_2*

such that the generalized UMD property (8.24) holds with

$$\mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \leq K_1 \log(Bn) \cdot \mathbb{E}_{\epsilon,\epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1})) + K_2.$$

The next example is that of homogenous polynomial classes under an injective tensor norm. The full description of this setting is deferred to Section 8.7.

Example 19 (Homogeneous Polynomials)**.** *Consider homogeneous polynomials of degree $2k$, with coefficients under the unit ball of the norm $(\|\cdot\|_{\{1,\dots,k\},\{k+1,\dots,2k\}})_\star$ in $(\mathbb{R}^d)^{\otimes 2k}$. Then there exist constants K_1, K_2 such that the generalized UMD property (8.24) holds with*

$$\mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \leq K_1 k^2 \log^2(d) \log(Bn) \cdot \mathbb{E}_{\epsilon,\epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1})) + K_2 k^2 \log^2(d).$$

8.6.2 Necessary Versus Sufficient Conditions

When we take $\mathcal{F}$ to be the unit ball of the dual norm $\|\cdot\|_\star$ as in previous sections, the inequality in (8.24) becomes:

$$\mathbb{E}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1}) \right\| \leq C \mathbb{E}_{\epsilon,\epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1}) \right\|. \tag{8.25}$$

This condition is sometimes referred to as a *probabilistic one-sided UMD inequality* for Paley-Walsh martingales (Hytönen et al., 2016). Comparing the condition to the UMD_1 inequality (8.9) one observes three differences: The Rademacher sequence ϵ' is drawn uniformly rather than being fixed, we only consider Paley-Walsh martingales (trees), and there is no supremum over end times. The supremum in (8.9) does not present a significant difference, as it can be removed from UMD_1 at a multiplicative cost of $O(\log n)$. The randomization over ϵ' is more interesting. It turns out that if in addition to (8.25) we require the opposite direction of this inequality to hold, i.e.

$$\mathbb{E}_{\epsilon,\epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1}) \right\| \leq C' \mathbb{E}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1}) \right\|,$$

then this is equivalent to the full UMD property (8.9) up to the presence of the supremum (Hytönen et al., 2016, Theorem 4.2.5). Thus, (8.25) can be thought of as a *one-sided* version of the UMD inequality.

There are indeed classes for which one-sided UMD inequality holds but the full UMD property does not. A result due to Hitczenko (1994) shows that there is a mild separation between these conditions even in the scalar setting:[6]

Theorem 13 (Hitczenko (1994)). *There exists a constant K independent of p such that for all $p \in [1, \infty)$,*

$$\mathop{\mathbb{E}}_{\epsilon}\left|\sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1})\right|^p \leq K^p \mathop{\mathbb{E}}_{\epsilon,\epsilon'}\left|\sum_{t=1}^{n} \epsilon_t' \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1})\right|^p. \tag{8.26}$$

When $p = 1$ this result is exactly the generalized UMD inequality (8.24), and for $p > 1$ it gives a one-sided version of the UMD_p condition. This bound is quantitatively stronger than what one would obtain from the UMD_p property, since (Burkholder, 1984) shows that the full two-sided UMD_p condition requires $K \geq p^\star - 1$. However, we remark that the gap here is only in logarithmic factors, and that the separation between the one-sided and full UMD properties is very mild for all examples we are aware of.

8.6.3 Application: Empirical Covering Number Bounds

Having developed online learning algorithms for which regret is bounded by the empirical Rademacher complexity, we are in the appealing position of being able to apply empirical process tools designed for the *statistical setting* to derive tight regret bounds for the *adversarial setting*. A powerful tool to derive instance-dependent upper bounds on the empirical Rademacher complexity is chaining.

[6] See also Hitczenko (1993); Cox and Veraar (2007, 2011).

Definition 5 (Empirical Cover). *For a hypothesis class $\mathcal{F} : \mathcal{X} \to \mathbb{R}$, data sequence $x_{1:n}$, and $\alpha > 0$, a set $\mathcal{V} \subseteq \mathbb{R}^n$ is called an empirical covering with respect to ℓ_p, $p \in [1, \infty)$, if*

$$\forall f \in \mathcal{F}\ \exists v \in \mathcal{V} \ \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{n}(f(x_t) - v_t)^p\right)^{1/p} \leq \alpha. \tag{8.27}$$

The set $\mathcal{V}$ is a cover with respect to ℓ_∞ if $\forall f \in \mathcal{F}\ \exists v \in \mathcal{V}$ s.t. $|f(x_t) - v_t| \leq \alpha\ \forall t \in [n]$.

We let the *empirical covering number* $\mathcal{N}_p(\mathcal{F}, \alpha, x_{1:n})$ denote the size of the smallest α-empirical cover for $\mathcal{F}$ on $x_{1:n}$ with respect to ℓ_p.

Because our task is simply to obtain bounds on the empirical Rademacher complexity on a particular sequence $x_{1:n}$, we can obtain regret bounds that depend on the data-dependent *empirical* covering number defined above, instead of a *worst-case* covering number. Such bounds have proved elusive in the adversarial setting, where most existing results are based on worst-case covering numbers (e.g. Rakhlin et al. (2010)). In particular, we derive two regret bounds based on the classical covering number bound (Pollard, 1990) and Dudley Entropy Integral bound (Dudley, 1967) for Rademacher complexity.

Theorem 14 (Empirical covering bound). *For any class $\mathcal{F} \subseteq [-1, +1]^{\mathcal{X}}$ satisfying the generalized* UMD *inequality (8.24) with constant C, there exists a strategy $(\widehat{y}_t)$ that attains*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq O\Big(C \cdot \inf_{\alpha > 0}\Big\{\alpha n + \sqrt{\log \mathcal{N}_1(\mathcal{F}, \alpha, x_{1:n}) n}\Big\}\Big). \tag{8.28}$$

Theorem 15 (Empirical Dudley Entropy bound). *For any class $\mathcal{F} \subseteq [-1, +1]^{\mathcal{X}}$ satisfying the generalized* UMD *inequality (8.24) with constant C, there exists a strategy $(\widehat{y}_t)$ that attains*

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq O\left(C \cdot \inf_{\alpha > 0}\left\{\alpha \cdot n + \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \delta, x_{1:n}) n} d\delta\right\}\right). \tag{8.29}$$

More generally, since our upper bounds depend on the empirical Rademacher complexity conditioned on the data $x_{1:n}$, more powerful techniques — such as Talagrand's generic chaining — may be applied to derive even tighter data-dependent covering bounds than those implied by (8.29).

8.7 Detailed Proofs and UMD Tools

8.7.1 Detailed Proofs

Proof of Lemma 13. Recall that $\ell_{\text{hinge}}(\widehat{y}, y) = \max\{0, 1 - \widehat{y} \cdot y\}$, $\ell_{\text{abs}}(\widehat{y}, y) = |\widehat{y} - y|$, $\ell_{\text{lin}}(\widehat{y}, y) = -\widehat{y} \cdot y$. Fix a sequence $x_{1:n}$, and let $y_t = \epsilon_t$ where $\epsilon \in \{\pm 1\}^n$ is a Rademacher sequence. By our hypothesis, we have

$$\mathcal{B}(x_{1:n}) \geq \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] \geq \mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right].$$

For the linear loss, observe that since $\widehat{y}_t$ cannot react to ϵ_t, we immediately have

$$\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] = \mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] = \widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}).$$

For the absolute and hinge losses, we will use two facts. First, since $|f(x_t)| \leq 1$, both losses satisfy $\ell(f(x_t), \epsilon_t) = 1 - f(x_t)\epsilon_t$. Second, without any assumption on the range of $\widehat{y}_t$, one has $\ell(\widehat{y}_t, \epsilon_t) \geq 1 - \widehat{y}_t\epsilon_t$. Therefore, whenever ℓ is the absolute or hinge loss, one has

$$\begin{aligned}
\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), \epsilon_t)\right] &\geq \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} (1 - \widehat{y}_t\epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} (1 - f(x_t)\epsilon_t)\right] \\
&= \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n} -\widehat{y}_t\epsilon_t - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} -f(x_t)\epsilon_t\right] \\
&= \mathop{\mathbb{E}}_{\epsilon}\left[-\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} -\epsilon_t f(x_t)\right].
\end{aligned}$$

The above is equal to $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n})$ as in the linear loss case, so we have shown that for each loss our hypothesis implies $\widehat{\mathcal{R}}(\mathcal{F}, x_{1:n}) \leq \mathcal{B}(x_{1:n})$. $\square$

Proof of Proposition 12. See proof of Theorem 11. □

8.7.2 Proofs from Section 8.4

Proof of Theorem 8. For the case $p, q \in (1, \infty)$, we appeal to Theorem 17.

Now consider the case $q = 1$, and suppose UMD_p holds for $p \in (1, \infty)$ with $\mathbf{C}_p$. Then by Theorem 17, $\mathbf{C}_2 \leq 200\mathbf{C}_p$. Finally, by Theorem 18, $\mathbf{C}_1 \leq 108\mathbf{C}_2 \leq 108 \cdot 200\mathbf{C}_p$.

For the converse direction, we appeal to Pisier (2011), Remark 8.2.4.

□

Proof of Theorem 9. Fix some $C > 0$ to be chosen later. Define the minimax value for the a game where the learner's goal is to achieve the adaptive regret bound:

$$\mathcal{V}_n^{\mathrm{ol}} =$$
$$\left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta_{[-B,+B]}} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t \right\| \right],$$

where we have adopted the shorthand $\mathcal{V}_n^{\mathrm{ol}} := \mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}, \mathcal{B})$. As is routine by now, there always exists some randomized strategy making predictions in $[-B, +B]$ whose regret is bounded by

$$C \,\mathbb{E} \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t \right\| + \mathcal{V}_n^{\mathrm{ol}}.$$

We will show that for the value of C given in the theorem statement one has $\mathcal{V}_n^{\mathrm{ol}} \leq 0$. To begin, observe that in view of the linearization inequality (8.6), the minimax value $\mathcal{V}_n^{\mathrm{ol}}$ is bounded by

$$\left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta_{[-B,+B]}} \sup_{y_t \in \mathcal{Y}} \mathop{\mathbb{E}}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell'(\widehat{y}_t, y_t)\widehat{y}_t + \left\| \sum_{t=1}^{n} \ell'(\widehat{y}_t, y_t) x_t \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t \right\| \right].$$

Using the minimax theorem as in Section 2.6,[7] the last expression is equal to

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta(\mathcal{Y})} \inf_{\widehat{y}_t \in [-B,+B]} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell'(\widehat{y}_t, y_t)\widehat{y}_t + \left\| \sum_{t=1}^{n} \ell'(\widehat{y}_t, y_t) x_t \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t \right\| \right].$$

Choose $\widehat{y}_t^\star = \arg\min_f \mathbb{E}_{y_t \sim p_t}[\ell(f, y_t)]$. By the assumption on the loss, the minimizer is obtained in $[-B, B]$ and so $\mathbb{E}_{y_t \sim p_t}[\ell'(\widehat{y}_t^\star, y_t)] = 0$. With this (sub)optimal choice, we obtain an upper bound of

$$\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell'(\widehat{y}_t^\star, y_t)\widehat{y}_t^\star + \left\| \sum_{t=1}^{n} \ell'(\widehat{y}_t^\star, y_t) x_t \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^\star, y_t) x_t \right\| \right].$$

Since $\widehat{y}_t^\star$ is the population minimizer, we have $\mathbb{E}_{y_t \sim p_t}[\ell'(\widehat{y}_t^\star, y_t)\widehat{y}_t^\star] = \mathbb{E}_{y_t \sim p_t}[\ell'(\widehat{y}_t^\star, y_t)]\widehat{y}_t^\star = 0$. The preceeding expression is thus equal to

$$\begin{aligned} &\left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\left\| \sum_{t=1}^{n} \ell'(\widehat{y}_t^\star, y_t) x_t \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^\star, y_t) x_t \right\| \right] \\ &\le \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \ell'(\widehat{y}_t^\star, y_t) x_t \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^\star, y_t) x_t \right\| \right]. \end{aligned}$$

Observe that we may rewrite the above expression as

$$\sup_{\mathbf{x}} \sup_{P} \mathop{\mathbb{E}}_{y_{1:n} \sim P} \left[\sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \ell'(\widehat{y}_t^\star(p_{1:t}), y_t) \mathbf{x}_t(y_{1:t-1}) \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^\star(p_{1:t}), y_t) \mathbf{x}_t(y_{1:t-1}) \right\| \right],$$

where $P = (p_1, \ldots, p_n)$ is a sequence of conditional distributions over $y_{1:n}$, $\mathbf{x}$ is a sequence of mappings $\mathbf{x}_t : \mathcal{Y}^{t-1} \to \mathcal{X}$, and $\widehat{y}_t^\star(p_{1:t})$ is the minimizer policy described above. For any fixed choice for P and $\mathbf{x}$, we have that $(\ell'(\widehat{y}_t^\star(p_{1:t}), y_t)\mathbf{x}_t(y_{1:t-1}))_{t \le n}$ is a martingale difference sequence, because the choice of $\widehat{y}_t^\star$ guarantees $\mathbb{E}[\ell'(\widehat{y}_t^\star(p_{1:t}), y_t)\mathbf{x}_t(y_{1:t-1}) \mid y_{1:t-1}] = 0$.

Therefore, if UMD_1 holds with constant $\mathbf{C}_1$, we have (by choosing a uniform random sign sequence in Definition 4) that for any fixed P, $\mathbf{x}$,

$$\mathbb{E} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \ell'(\widehat{y}_t^\star(p_{1:t}), y_t) \mathbf{x}_t(y_{1:t-1}) \right\| \le \mathbf{C}_1 \, \mathbb{E} \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \le n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^\star(p_{1:t}), y_t) \mathbf{x}_t(y_{1:t-1}) \right\|.$$

[7] A word of caution: we use the assumption on the loss that there exists a minimizer for every label within some bounded domain precisely so that we can now use minimax theorem restricting $\hat{y}_t$'s to be in bounded domain.

This implies that the inequality holds for the supremum over P and $\mathbf{x}$, so we have

$$\mathcal{V}_n^{\text{ol}} \leq \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta(\mathcal{Y})} \mathop{\mathbb{E}}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\mathbf{C}_1 \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^{\star}, y_t) x_t \right\| - C \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t^{\star}, y_t) x_t \right\| \right].$$

Thus, if we take $C \geq \mathbf{C}_1$:

$$\leq 0.$$

We have established that there exists a strategy $(\widehat{y}_t)$ guaranteeing

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \leq \mathbf{C}_1 \, \mathbb{E} \mathop{\mathbb{E}}_{\epsilon} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t \right\|$$

Treating $(\ell'(\widehat{y}_t, y_t) x_t)_{t \leq n}$ as a fixed sequence, we may now apply Corollary 10 to remove the supremum over end times:

$$\leq 2\mathbf{C}_1 \log\left(\max_{t \in [n]} \|x_t\| n \right) \cdot \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t \right\| + 2\mathbf{C}_1.$$

By the standard contraction argument for Rademacher complexity, since $|\ell'| \leq 1$,

$$\leq 2\mathbf{C}_1 \log\left(\max_{t \in [n]} \|x_t\| n \right) \cdot \mathop{\mathbb{E}}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\| + 2\mathbf{C}_1.$$

Finally, recall that by Theorem 8, $\mathbf{C}_1 \leq O(\mathbf{C}_p)$.

□

Proof of Theorem 10. Most of the proofs in this theorem use the following fact: If $(X_t)_{t \leq n}$ is a martingale difference sequence, its restriction to a subset of coordinates is also a martingale difference sequence. This allows one to prove the deterministic UMD property (8.8) for complex spaces by building up from simpler spaces.

- $(\mathbb{R}, |\cdot|)$: Burkholder (1984) shows that for all $p \in (1, \infty)$, $\mathbf{C}_p = p^{\star} - 1$.

- $(\mathbb{R}^d, \|\cdot\|_p)$, for $p \in (1, \infty)$:

$$\mathop{\mathbb{E}}_{X}\left\|\sum_{t=1}^{n} \epsilon_t X_t\right\|_p^p = \sum_{i\in[d]} \mathop{\mathbb{E}}_{X}\left|\sum_{t=1}^{n} \epsilon_t X_t[i]\right|^p \leq (p^\star - 1) \sum_{i\in[d]} \mathop{\mathbb{E}}_{X}\left|\sum_{t=1}^{n} X_t[i]\right|^p = (p^\star - 1) \mathop{\mathbb{E}}_{X}\left\|\sum_{t=1}^{n} X_t\right\|_p^p. \tag{8.30}$$

  The middle inequality here uses the UMD_p constant for the scalar case.

- $(\mathbb{R}^d, \|\cdot\|_p)$, for $p \in \{1, \infty\}$: We will start with ℓ_∞. Set $p = \log d$, and observe that for ℓ_p, by Theorem 17, ℓ_p has $\mathbf{C}_2 = O(\mathbf{C}_p) = O(p^\star)$ (the second bound is from the previous example). Then we have, for any sequence of signs,

$$\begin{aligned}
\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t X_t\right\|_\infty^2 &\leq \mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t X_t\right\|_p^2 \\
&\leq O(p^\star)\,\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|_p^2 \\
&\leq O(p^\star)\,\mathbb{E}\left(d^{1/p}\left\|\sum_{t=1}^{n} X_t\right\|_\infty\right)^2.
\end{aligned}$$

  Since $d^{1/\log d} = O(1)$, the last expression is at most

$$O(p^\star)\,\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|_\infty^2.$$

  Finally, note that $p^\star = O(\log d)$.

  The same argument works for the ℓ_1 norm using $p = 1 + 1/\log d$. Alternatively, the constant can be deduced from duality using Theorem 19. That these constants are optimal follows from Hytönen et al. (2016), Proposition 4.2.19.

- $(\mathbb{R}^d, \|\cdot\|_{\mathcal{A}}/\|\cdot\|_{\mathcal{A}^\star})$. Let us focus on $\|\cdot\|_{\mathcal{A}^\star}$. Assume $\mathcal{A} = \{a_1, \ldots, a_N\}$. Observe that

$$\begin{aligned}
\|x\|_{\mathcal{A}^\star} &= \max\{\langle y, x\rangle \mid y \in \text{conv}(\mathcal{A})\} \\
&= \max\left\{\sum_{i\in[N]} \theta_i \langle a_i, x_i\rangle \mid \theta \in \Delta(N)\right\}
\end{aligned}$$

Since we assumed $\mathcal{A}$ is symmetric:

$$= \left\| (\langle a_i, x_i \rangle)_{i\in[N]} \right\|_\infty$$

$= \|Ax\|_\infty$, where $A \in \mathbb{R}^{N\times d}$ is the matrix of elements of $\mathcal{A}$ stacked as rows.

For any martingale difference sequence $(X_t)_{t\leq n}$, $(AX_t)_{t\leq n}$ is also a martingale difference. Therefore, we can deduce the UMD_2 property for $\|\cdot\|_{\mathcal{A}^\star}$ from our result for $\|\cdot\|_\infty$. The UMD_2 property for $\|\cdot\|_{\mathcal{A}}$ follows from Theorem 19.

- $(\mathbb{R}^{d\times d}, \|\cdot\|_{S_p})$, for $p \in (1,\infty)$: Hytönen et al. (2016) Theorem 5.2.10 and Proposition 5.5.5.

- $(\mathbb{R}^{d\times d}, \|\cdot\|_\sigma)$: $\mathbf{C}_2 = O(\log^2 d)$. We will build up from the Schatten p-norms in the same fashion as for the ℓ_p spaces. Let $p = \log d$. For any sequence of signs,

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_\sigma^2 \leq \mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_{S_p}^2.$$

Using Theorem 17 to get $C_2 \leq O((p^\star)^2)$ for S_p:

$$\leq O((p^\star)^2)\,\mathbb{E}\left\|\sum_{t=1}^n X_t\right\|_{S_p}^2$$

$$\leq O((p^\star)^2)\,\mathbb{E}\left(d^{1/p}\left\|\sum_{t=1}^n X_t\right\|_\sigma\right)^2.$$

Since $d^{1/\log d} = O(1)$, the preceding expression is at most

$$O((p^\star)^2)\,\mathbb{E}\left\|\sum_{t=1}^n X_t\right\|_\sigma^2.$$

Once again, $p^\star \leq \log d$. The constant for $\|\cdot\|_\Sigma$ follows from Theorem 19, since the trace norm is dual to the spectral norm.

- $(\mathbb{R}^{d\times d}, \|\cdot\|_{p,q})$, for $p, q \in (1,\infty)$: For any sequence of signs, we apply the UMD property for ℓ_p row-wise:

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t X_t\right\|_{p,q}^p = \sum_{i\in[d]} \mathbb{E}\left\|\sum_{t=1}^n \epsilon_t (X_t)_{i\cdot}\right\|_q^p.$$

We know ℓ_q has $\mathbf{C}_q \leq O(q^\star)$. By Theorem 17, this implies that $\mathbf{C}_p$ for ℓ_q has $\mathbf{C}_p \leq O(p^\star \cdot q^\star)$.

$$\leq O(p^\star \cdot q^\star) \sum_{i\in[d]} \mathbb{E}\left\|\sum_{t=1}^{n} (X_t)_{i\cdot}\right\|_q^p$$
$$= O(p^\star \cdot q^\star)\, \mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|_{p,q}^p.$$

- $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ for any Hilbert space $\mathcal{H}$: See Example 15.

□

8.7.3 Proofs from Section 8.5

Proof of Theorem 11. We will show that the strategy achieves the regret bound

$$\mathbb{E}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'(\widehat{y}_t^{\epsilon_{1:t-1}}, y_t) x_t\right\|^p\right)\right] \leq 0. \quad (8.31)$$

Our proof follows the same step-by-step minimax analysis as variants of the Burkholder algorithm from previous chapters.

Initial Condition The first step is to show that the Burkholder function upper bounds that difference between regret and the desired regret bound. In view of (8.6),

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'(\widehat{y}_t, y_t) x_t\right\|^p\right)$$
$$\leq \sum_{t=1}^{n} \widehat{y}_t \ell'_t + \left\|\sum_{t=1}^{n} \ell'_t x_t\right\| - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right)$$

$$\leq \sum_{t=1}^{n} \widehat{y}_t \ell'_t + \Psi_{\eta,p}\left(\left\|\sum_{t=1}^{n} \ell'_t x_t\right\|^p\right) - \Psi_{\eta,p}\left(\beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right)$$
$$= \sum_{t=1}^{n} \widehat{y}_t \ell'_t + \frac{\eta}{p}\left(\left\|\sum_{t=1}^{n} \ell'_t x_t\right\|^p - \beta^p \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|^p\right)$$
$$\leq \sum_{t=1}^{n} \widehat{y}_t \ell'_t + \frac{\eta}{p}\mathbf{U}_p\left(\sum_{t=1}^{n} \ell'_t x_t, \sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right).$$

Admissibility Condition At each time step $t \in [n]$, we have the following recursive upper bound on the cost to go

$$\sup_{x_t} \inf_{\widehat{y}_t} \sup_{\ell'_t} \mathbb{E}_{\epsilon_t}\left[\widehat{y}_t \ell'_t + \frac{\eta}{p}\mathbf{U}_p\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{t=1}^{t} \epsilon_s \ell'_s x_s\right)\right]$$
$$= \sup_{x_t} \inf_{\widehat{y}_t} \sup_{\ell'_t}\left[\widehat{y}_t \ell'_t + \mathbb{E}_{\epsilon_t} \frac{\eta}{p}\mathbf{U}_p\left(\sum_{s=1}^{t} \ell'_s x_s, \sum_{t=1}^{t} \epsilon_s \ell'_s x_s\right)\right]$$
$$= \sup_{x_t} \inf_{\widehat{y}_t} \sup_{\ell'_t}[\widehat{y}_t \ell'_t + G_t(\ell'_t)].$$

Plugging in the strategy specified by Algorithm 5, the last expression is at most

$$\sup_{x_t} \sup_{\ell'_t}[-G'_t(0) \cdot \ell'_t + G_t(\ell'_t)] \leq \sup_{x_t} G_t(0) = \mathbf{U}_p\left(\sum_{s=1}^{t-1} \ell'_s x_s, \sum_{t=1}^{t-1} \epsilon_s \ell'_s x_s\right).$$

Finally, we have $\mathbf{U}_p(0,0) \leq 0$, and so the final value of the game is at most zero. This implies that (8.31) is achieved. $\square$

Proof of Lemma 14. In what follows we will leave the dependence of $\widehat{y}_t, x_t, \ell'_t$ on $\epsilon_{1:t-1}$ implicit for notational convenience. We will handle this dependence at the end of the proof. Assume $N > 1$. Otherwise, the algorithm's regret is bounded as $2\eta_1^{-(p'-1)} = 4\eta_0^{-(p'-1)}$. We begin with the elementary upper bound

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^{n}\ell(\widehat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t), y_t)\right] \\
&\leq \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{i=1}^{N}\left[\sum_{t=s_i}^{s_{i+1}-1}\ell(\widehat{y}_t^{\epsilon_{1:t-1}}, y_t) - \inf_{f\in\mathcal{F}}\sum_{t=s_i}^{s_{i+1}-1}\ell(f(x_t), y_t)\right]\right].
\end{aligned}$$

Using the regret bound for Algorithm 5 (note that that algorithm has an anytime regret guarantee) given by Theorem 11:

$$\leq \mathop{\mathbb{E}}_{\epsilon}\left[\frac{1}{p}\sum_{i=1}^{N}\left[\eta_i\beta_p^p\left\|\sum_{t=s_i}^{s_{i+1}-1}\epsilon_t\ell'_t x_t\right\|^p + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]\right].$$

Introducing a new supremum:

$$\leq \mathop{\mathbb{E}}_{\epsilon}\left[\frac{1}{p}\sum_{i=1}^{N}\left[\eta_i\Phi(x_{s_i:s_{i+1}-1}, \ell'_{s_i:s_{i+1}-1}, \epsilon_{s_i:s_{i+1}-1}) + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]\right].$$

The doubling condition implies that $\eta_i\Phi(x_{s_i:s_{i+1}-2}, \ell'_{s_i:s_{i+1}-2}, \epsilon_{s_i:s_{i+1}-2}) \leq \eta_i^{-(p'-1)}$. To use this fact, observe that since $\|x_t\| \leq 1$, we have that for any $C > 0$,

$$\begin{aligned}
&\eta_i\Phi(x_{s_i:s_{i+1}-1}, \ell'_{s_i:s_{i+1}-1}, \epsilon_{s_i:s_{i+1}-1}) \\
&= \eta_i\beta_p^p \sup_{s_i\leq a\leq b\leq s_{i+1}-1}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|^p \\
&\leq \eta_i(1+1/C)^p\beta_p^p \sup_{s_i\leq a\leq b\leq s_{i+1}-2}\left\|\sum_{t=a}^{b}\epsilon_t\ell'_t x_t\right\|^p + \eta_i C^p\beta_p^p.
\end{aligned}$$

For $C = p$:

$$\begin{aligned}
&\leq \eta_i e\Phi(x_{s_i:s_{i+1}-2}, \epsilon_{s_i:s_{i+1}-2}) + \eta_i p^p\beta_p^p. \\
&= e\eta_i^{-(p'-1)} + \eta_i p^p\beta_p^p.
\end{aligned}$$

Returning to the regret bound, we have

$$\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\epsilon}\left[\frac{1}{p}\sum_{i=1}^{N}\left[e\eta_i^{-(p'-1)} + \eta_i p^p\beta_p^p + \frac{1}{p'-1}\eta_i^{-(p'-1)}\right]\right] \\
&\leq \mathop{\mathbb{E}}_{\epsilon}\left[e\sum_{i=1}^{N}\eta_i^{-(p'-1)} + p^p\beta_p^p\eta_i\right]
\end{aligned}$$

We will handle with the left-hand term first. Now observe that

$$\eta_{N-1}\Phi(x_{s_{N-1}:s_N}, \ell'_{s_{N-1}:s_N}, \epsilon_{s_{N-1}:s_N}) > \eta_{N-1}^{-(p'-1)}.$$

Rearranging further implies

$$\eta_{N-1}^{-(p'-1)} \leq \Phi(x_{s_{N-1}:s_N}, \ell'_{s_{N-1}:s_N}, \epsilon_{s_{N-1}:s_N})^{1/p} \leq \Phi(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n})^{1/p}.$$

Finally, since $\eta_i^{-(p'-1)} = 2\eta_{i-1}^{-(p'-1)}$,

$$\sum_{i=1}^{N} \eta_i^{-(p'-1)} = \eta_0^{-(p'-1)} \sum_{i=1}^{N} 2^i \leq 2 \cdot 2^N \eta_0^{-(p'-1)} \leq 4\Phi(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n})^{1/p} = 4\beta_p \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\|.$$

For the second term, observe that $\eta_i \leq \eta_0$ for all i, so

$$\sum_{i=1}^{N} p^p \beta_p^p \eta_i \leq p^p \beta_p^p \eta_0 \cdot N.$$

Finally, by the invariant $2^{N-1}\eta_0^{-(p'-1)} \leq \Phi(x_{1:n}, \epsilon_{1:n})^{1/p}$ we established earlier,

$$N \leq \log\left(\Phi(x_{1:n}, \ell'_{1:n}, \epsilon_{1:n})^{1/p} \eta_0^{(p'-1)}\right) + 1$$

Putting everything together, the regret is bounded as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon} \max\left\{ 2e\beta_p \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| + p^p \beta_p^p \eta_0 \left(\log\left(\sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| \eta_0^{(p'-1)} \right) + 1 \right), 4\eta_0^{-(p'-1)} \right\} \\
&\leq \mathop{\mathbb{E}}_{\epsilon} \left[2e\beta_p \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| + p^p \beta_p^p \eta_0 \left(\log\left(\sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| \eta_0^{(p'-1)} \right) + 1 \right) + 4\eta_0^{-(p'-1)} \right]
\end{aligned}$$

Using that $\|x_t\| \leq 1$:

$$\leq 2e\beta_p \mathop{\mathbb{E}}_{\epsilon} \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| + p^p \beta_p^p \eta_0 \log\left(n \cdot \eta_0^{(p'-1)}\right) + 4\eta_0^{-(p'-1)}.$$

For the choice $\eta_0 = (\beta_p \cdot p)^{-p}$:

$$\leq 2e\beta_p \mathop{\mathbb{E}}_{\epsilon} \sup_{1\leq a\leq b\leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| + \log(n) + (p \cdot \beta_p)^{\frac{p}{p-1}}.$$

For the choice $\eta_0 = 1$:

$$\leq 2e\beta_p \mathop{\mathbb{E}}_{\epsilon} \sup_{1 \leq a \leq b \leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| + p^p \beta_p^p \log(n) + 4.$$

Writing $x_t(\epsilon_{1:t-1})$ and $\ell'_t(\epsilon_{1:t-1})$ to make the adversary's dependence on the sequence ϵ explicit, the main term of interest in the above quantity is

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{1 \leq a \leq b \leq n} \left\| \sum_{t=a}^{b} \epsilon_t \ell'_t(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\|.$$

It remains to remove the supremum and decouple the data sequences (x_t) and (ℓ'_t) from the Rademacher sequence ϵ. Since $\ell'_t x_t$ can only react to $\epsilon_{1:t-1}$, the sequence $(\epsilon_t \ell'_t x_t)_{t \leq n}$ is a martingale difference sequence. Since $\left\| \sum_{t=a}^{b} \epsilon_t \ell'_t x_t \right\| \leq n$, we may apply Corollary 9 to arrive at an upper bound of

$$\leq O\left(\log(n) \mathop{\mathbb{E}}_{\epsilon} \sup_{1 \leq b \leq n} \left\| \sum_{t=1}^{b} \epsilon_t \ell'_t(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\| \right).$$

Now observe that since Algorithm 5 uses a Burkholder function $\mathbf{U}_p$ for $(\|\cdot\|, p, \beta_p)$, Theorem 7 and Theorem 8 together imply that the UMD_1 inequality (8.9) holds with constant $O(\beta_p)$, therefore, the above is bounded as

$$\leq O\left(\beta_p \log(n) \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{\epsilon'} \sup_{1 \leq b \leq n} \left\| \sum_{t=1}^{b} \epsilon'_t \ell'_t(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\| \right).$$

Note that the variables (x_t) and (ℓ'_t) no longer depend on the Rademacher sequence appearing in the sum. Lastly, we apply Corollary 9 once more to remove the remaining supremum and arrive at the bound,

$$\leq O\left(\beta_p \log^2(n) \mathop{\mathbb{E}}_{\epsilon} \mathop{\mathbb{E}}_{\epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \ell'_t(\epsilon_{1:t-1}) x_t(\epsilon_{1:t-1}) \right\| \right).$$

□

Proof of Example 17. (8.21) is obtained by plugging the optimal UMD constant $p^\star - 1$ into the bound for Lemma 14. For (8.22), observe that for any sequence z_t we have $\mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t z_t\|_2 \leq \sqrt{\mathbb{E}_\epsilon \|\sum_{t=1}^n \epsilon_t z_t\|_2^2} = \sqrt{\mathbb{E}_\epsilon \sum_{t=1}^n \|z_t\|_2^2}$. Applying this fact with the algorithm's bound for $p = 2$ gives the regret bound

$$O\left(\sqrt{\sum_{t=1}^n \|\ell_t' x_t\|_2^2} \cdot \log^2 n + \log n\right).$$

For (8.23), observe that with $p = 1/\log d$ we have the regret bound

$$O\left(\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t \ell_t' x_t \right\|_p \cdot \log d \log^2 n + \log^2 d \log n\right).$$

However for any X, $\|X\|_p \leq d^{1-1/p} \|X\|_1$. For our choice of $p = 1 + 1/\log d$ we have $d^{1-1/p} = O(1)$.

$$\begin{aligned}
&\leq O\left(\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t \ell_t' x_t \right\|_1 \cdot \log d \log^2 n + \log^2 d \log n\right) \\
&\leq O\left(\mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t x_t \right\|_1 \cdot \log d \log^2 n + \log^2 d \log n\right) \\
&= O\left(\sum_{i \in [d]} \mathbb{E}_\epsilon \left| \sum_{t=1}^n \epsilon_t x_t[i] \right| \cdot \log d \log^2 n + \log^2 d \log n\right) \\
&\leq O\left(\sum_{i \in [d]} \sqrt{\sum_{t=1}^n (x_t[i])^2} \cdot \log d \log^2 n + \log^2 d \log n\right) \\
&= O\left(\sum_{i \in [d]} \|x_{1:n,i}\|_2 \cdot \log d \log^2 n + \log^2 d \log n\right).
\end{aligned}$$

□

8.7.4 Proofs from Section 8.6

Since we restrict to the absolute loss in this section and restrict to $y_t \in [-1, +1]$, we can also restrict to $\widehat{y}_t \in [-1, +1]$ without loss of generality, since for any value of y_t the loss may always be decreased by clipping $\widehat{y}_t$ into this range. In the proof below, any infimum over $\widehat{y}_t$ is understood to be over this range.

Proof of Theorem 12. We shall first show that condition 2 implies condition 1, specifically for constant $B = 2C$. We can write down the minimax value for the proposed regret bound and check if it indeed is achievable. To this end, note that

$$
\begin{aligned}
&\mathcal{V}_n^{\mathrm{ol}} \\
&= \left\langle\!\!\left\langle \sup_{x_t} \inf_{\widehat{y}_t} \sup_{y_t \in [-1,+1]} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\ell(\widehat{y}_t, y_t) - \ell(f(x_t), y_t)) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&\le \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \ell'(\widehat{y}_t, y_t)(\widehat{y}_t - f(x_t)) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right]
\end{aligned}
$$

Setting $\widehat{y}_t^*$ to be minimizer of $\mathbb{E}\, \ell(\hat{y}_t, y_t)$, we have

$$
\begin{aligned}
&\le \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \ell'(\widehat{y}_t^*, y_t)(\widehat{y}_t^* - f(x_t)) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \inf_{\widehat{y}_t} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} -\ell'(\widehat{y}_t^*, y_t) f(x_t) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \left(\mathbb{E}_{y_t' \sim p_t} \ell'(\widehat{y}_t^*, y_t') - \ell'(\widehat{y}_t^*, y_t) \right) f(x_t) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&\le \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathbb{E}_{y_t, y_t' \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} (\ell'(\widehat{y}_t^*, y_t') - \ell'(\widehat{y}_t^*, y_t)) f(x_t) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{x_t} \sup_{p_t \in \Delta[-1,+1]} \mathbb{E}_{y_t, y_t' \sim p_t} \mathbb{E}_{\epsilon_t'} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} \epsilon_t'(\ell'(\widehat{y}_t^*, y_t') - \ell'(\widehat{y}_t^*, y_t)) f(x_t) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&\le \left\langle\!\!\left\langle \sup_{x_t} \mathbb{E}_{\epsilon_t'} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} 2\epsilon_t' f(x_t) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \\
&= \sup_{\mathbf{x}} \mathbb{E}_{\epsilon'} \sup_{f \in \mathcal{F}} \left[\sum_{t=1}^{n} 2\epsilon_t' f(\mathbf{x}_t(\epsilon_{1:t-1}')) - 2C\, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon_{1:t-1}') \right].
\end{aligned}
$$

However, by condition 2, we have that the above is bounded by 0 and so we can conclude that the minimax strategy does attain the regret bound proposed in condition 1.

Now to prove that condition 1 implies condition 2 (with constant B), notice that we have an algorithm that guarantees regret bound:

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq B \, \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t)$$

Assume now that the adversary at time first provides input instance $\mathbf{x}_t(\epsilon_{1:t-1})$ where $\mathbf{x}$ is any arbitrary $\mathcal{X}$ valued binary tree. Also assume that y_t is picked to be ϵ_t a draw of a coin flip. In this case, we have from the regret bound that

$$\sum_{t=1}^{n} \ell(\hat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t(\epsilon_{1:t-1})), \epsilon_t) \leq B \, \mathbb{E}_{\epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

Taking expectation we find that,

$$\mathbb{E}_{\epsilon}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, \epsilon_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t(\epsilon_{1:t-1})), \epsilon_t)\right] \leq B \, \mathbb{E}_{\epsilon, \epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

Now notice that irrespective of what $\hat{y}_t$ the algorithm picks, $\mathbb{E}_{\epsilon_t} \ell(\hat{y}_t, \epsilon_t) = 1$. Hence,

$$\mathbb{E}_{\epsilon}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} (1 - \ell(f(\mathbf{x}_t(\epsilon_{1:t-1})), \epsilon_t))\right] \leq B \, \mathbb{E}_{\epsilon, \epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

However note that when $y \in \{\pm 1\}$ and $a \in [-1, 1]$, we have that $\ell(a, y) = |a - y| = 1 - ay$. Hence from above we conclude that,

$$\mathbb{E}_{\epsilon}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1}))\right] \leq B \, \mathbb{E}_{\epsilon, \epsilon'} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1}))$$

Since the above is true for any choice of x, we have shown that condition 1 implies condition 2 with constant B. $\square$

Proof of Example 18. Let x be some $\mathcal{X}$-valued tree. Observe that by the reproducing property,

$$\mathbb{E}_{\sigma} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \sigma_t f(\mathbf{x}_t(\sigma)) = \mathbb{E}_{\sigma} \left\| \sum_{t=1}^{n} \sigma_t K(\cdot, \mathbf{x}_t(\sigma)) \right\|_{\mathcal{H}},$$

and likewise $\mathbb{E}_{\sigma,\epsilon}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\sigma)) = \mathbb{E}_{\sigma,\epsilon}\|\sum_{t=1}^{n}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\|_{\mathcal{H}}$.

Since $\mathcal{H}$ is a Hilbert space the deterministic UMD property for power 2 is trivial. For any fixed sequence $\epsilon \in \{\pm 1\}^n$,

$$\mathbb{E}_{\sigma}\left\|\sum_{t=1}^{n}\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}^2 = \mathbb{E}_{\sigma}\left\|\sum_{t=1}^{n}\epsilon_t\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}^2.$$

By Corollary 11, this implies there is some C such that

$$\mathbb{E}_{\sigma}\sup_{\tau\leq n}\left\|\sum_{t=1}^{\tau}\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}} = C\,\mathbb{E}_{\sigma}\sup_{\tau\leq n}\left\|\sum_{t=1}^{\tau}\epsilon_t\sigma_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}}.$$

Now suppose ϵ is drawn uniformly at random. For a fixed draw of σ, Corollary 10 implies that the RHS enjoys the bound

$$\mathbb{E}_{\epsilon}\sup_{\tau\leq n}\left\|\sum_{t=1}^{\tau}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}} \leq 2\log(Bn)\,\mathbb{E}_{\epsilon}\left\|\sum_{t=1}^{n}\epsilon_t K(\cdot,\mathbf{x}_t(\sigma))\right\|_{\mathcal{H}} + 2$$

$\square$

Polynomials

Suppose we receive data $x_1,\ldots,x_n \in \mathbb{R}^d$ and want to compete with a class $\mathcal{F}$ of homogeneous polynomials of degree k. Any homogeneous degree k polynomial f may be represented via a coefficient tensor M in $(\mathbb{R}^d)^{\otimes k}$ via

$$f(x) = \langle M, x^{\otimes k}\rangle.$$

We may take M to be symmetric, so that $M_{1,\ldots,k} = M_{\pi(1),\ldots,\pi(k)}$ for any permutation. We may thus work with a class $\mathcal{M} \subseteq (\mathbb{R}^d)^{\otimes k}$ of symmetric tensors, then take $\mathcal{F} = \{x \mapsto \langle M, x^{\otimes k}\rangle \mid M \in \mathcal{M}\}$. Our task is then to decide which norm to place on $\mathcal{M}$. Following, e.g., Adamczak and Wolff (2015); Wang et al. (2017), we define a class of general tensor norms. Let $\mathcal{J} = \{J_1, \ldots, J_N\}$ be a partition of $[k]$. For some $\alpha \in [d]^k$ and $J \subseteq [k]$, let $\alpha_J = (\alpha_i)_{i \in J}$. We then define

$$\|M\|_{\mathcal{J}} = \sup\left\{ \sum_{\alpha \in [d]^k} M_\alpha \prod_{l=1}^{N} x^l_{\alpha_{J_l}} \mid \left\|x^l\right\|_2 \leq 1 \; \forall l \in [N] \right\}, \tag{8.32}$$

where $x^l \in (\mathbb{R}^d)^{\otimes |J_l|}$. Under this notation we have $\|M\|_{\{1\},\{2\}}$ as the spectral norm and $\|M\|_{\{1,2\}}$ as the Frobenius norm when $k = 2$ and M is a matrix. In general, $\|M\|_{\{1\},\{2\},\ldots,\{k\}}$ is called the *injective tensor norm*.

Proof of Example 19. Fix an $\mathcal{X}$-valued tree x. Then we have

$$\mathop{\mathbb{E}}_{\sigma} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \sigma_t f(\mathbf{x}_t(\sigma)) = \mathop{\mathbb{E}}_{\sigma} \sup_{M \in \mathcal{M}} \sum_{t=1}^{n} \sigma_t \left\langle M, \mathbf{x}_t(\sigma)^{\otimes 2k} \right\rangle = \mathop{\mathbb{E}}_{\sigma} \left\| \sum_{t=1}^{n} \sigma_t \mathbf{x}_t(\sigma)^{\otimes 2k} \right\|_{\{1,\ldots,k\},\{k+1,\ldots,2k\}}$$

For some tensor $T \in (\mathbb{R}^d)^{\otimes 2k}$, we can define its flattening $\overline{T}$ into a $\mathbb{R}^{d^k \times d^k}$ matrix and verify that in fact

$$\|T\|_{\{1,\ldots,k\},\{k+1,\ldots,2k\}} = \max_{u,v \in \mathbb{R}^{d^k} \mid \|u\|_2, \|v\|_2 \leq 1} \sum_{\alpha \in [d]^k, \beta \in [d]^k} T_{\alpha,\beta} u_\alpha v_\beta = \left\langle u, \overline{T} v \right\rangle = \|\overline{T}\|_\sigma,$$

so in fact this is the spectral norm of the flattened matrix. Let $\mathbf{X}_t \in \mathbb{R}^{d^k \times d^k}$ be the flattening of $(\mathbf{x}_t)^{\otimes 2k}$. Then

$$\mathop{\mathbb{E}}_{\sigma} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \sigma_t f(\mathbf{x}_t(\sigma)) = \mathop{\mathbb{E}}_{\sigma} \left\| \sum_{t=1}^{n} \sigma_t \mathbf{X}_t(\sigma) \right\|_\sigma,$$

so we can prove the desired inequality by applying the UMD inequality for the spectral norm. Recall from Theorem 10 that the UMD inequality for the spectral norm has a constant of order $\log^2(\dim)$, which for this application translates into a constant of order $O(k^2 \log^2(d))$. We finally apply Corollary 10 as in Example 18 to get the result. $\square$

Empirical Covering Number Bounds

Proof of Theorem 14 and Theorem 15. Theorem 12 proves that when the one-sided UMD-property (8.24) holds, there exists a strategy whose regret is bounded as

$$C\,\mathbb{E}_{\epsilon}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t).$$

Since this quantity is the statistical Rademacher complexity, we may apply the classical covering number bound (Rakhlin and Sridharan, 2012, Proposition 12.3):

$$\mathbb{E}_{\epsilon}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t) \le O\Big(\inf_{\alpha>0}\Big\{\alpha n + \sqrt{\log\mathcal{N}_1(\Delta_d,\alpha,x_{1:n})n}\Big\}\Big).$$

Likewise, the classical Dudley entropy integral bound (Rakhlin and Sridharan, 2012, Theorem 12.4) yields:

$$\mathbb{E}_{\epsilon}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t) \le O\left(\inf_{\alpha>0}\left\{\alpha\cdot n + \int_{\alpha}^{1}\sqrt{\log\mathcal{N}_2(\mathcal{F},\delta,x_{1:n})n}d\delta\right\}\right).$$

□

8.7.5 UMD Spaces and Martingale Inequalities

Stopping Inequalities

Let (Z_t) be a martingale. For two stopping times τ_1,τ_2, we define its stopped version as $Z_t^{\tau_1:\tau_2}$ via

$$dZ_t^{\tau_1:\tau_2} = dZ_t\mathbb{1}\{t>\tau_1\}\mathbb{1}\{t\le\tau_2\}.$$

Proposition 13 (Hytönen et al. (2016), Proposition 3.1.14)**.** For any $p\in[1,\infty)$,

$$\mathbb{E}\|Z_n^{\tau_1:\tau_2}\|^p \le 2^p\,\mathbb{E}\|Z_n\|^p.$$

Theorem 16 (Doob's Maximal Inequality). *For any martingale $(Z_t)_{t\geq 1}$ taking values in $(\mathfrak{B}, \|\cdot\|)$ and any $p \in (1, \infty]$,*

$$\mathbb{E} \sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} dZ_t \right\|^p \leq (p')^p \, \mathbb{E} \left\| \sum_{t=1}^{n} dZ_t \right\|^p. \tag{8.33}$$

Furthermore

$$\mathbb{P}\left(\sup_{\tau \leq n} \left\| \sum_{t=1}^{\tau} dZ_t \right\| > \lambda \right) \leq \frac{1}{\lambda} \mathbb{E} \left\| \sum_{t=1}^{n} dZ_t \right\| \quad \forall \lambda > 0. \tag{8.34}$$

More generally, (8.33) and (8.34) hold when the sequence $(\|\sum_{t=1}^{\tau} dZ_t\|)_{\tau\geq 1}$ is replaced by any non-negative submartingale $(F_\tau)_{\tau \geq 1}$.

Corollary 9. If (F_n) is a non-negative submartingale and $F_n \leq A$ almost surely then for all $\eta > 0$,

$$\mathbb{E}\left[\max_{\tau \leq n} F_\tau \right] \leq (\log A + \log \eta) \cdot \mathbb{E}[F_n] + \frac{1}{\eta}.$$

Proof of Corollary 9.

$$\begin{aligned}
\mathbb{E}\left[\max_{\tau \leq n} F_\tau \right] &= \int_0^\infty \mathbb{P}\left(\max_{\tau \leq n} F_\tau > \lambda \right) d\lambda \\
&= \int_0^A \mathbb{P}\left(\max_{\tau \leq n} F_\tau > \lambda \right) d\lambda \\
&\leq \int_{1/\eta}^A \mathbb{P}\left(\max_{\tau \leq n} F_\tau > \lambda \right) d\lambda + \frac{1}{\eta} \\
&\leq \mathbb{E}[F_n] \int_{\frac{1}{\eta}}^A \frac{1}{\lambda} d\lambda + \frac{1}{\eta} \\
&= (\log A + \log \eta) \cdot \mathbb{E}[F_n] + \frac{1}{\eta}.
\end{aligned}$$

□

Corollary 10. Let (Z_t) be any martingale difference sequence in $(\mathfrak{B}, \|\cdot\|)$ with $\|Z_t\| \leq B$ almost surely. Then

$$\mathbb{E} \sup_{1 \leq a \leq b \leq n} \left\| \sum_{t=a}^{b} Z_t \right\| \leq 2 \log(Bn) \, \mathbb{E} \left\| \sum_{t=1}^{n} Z_t \right\| + 2. \tag{8.35}$$

Proof of Corollary 10. First, observe that the supremum over starting times can easily be removed:

$$\left\|\sum_{t=a}^{b} Z_t\right\| = \left\|\sum_{t=1}^{b} Z_t - \sum_{t=1}^{a-1} Z_t\right\| \leq \left\|\sum_{t=1}^{b} Z_t\right\| + \left\|\sum_{t=1}^{a-1} Z_t\right\|,$$

and so

$$\mathbb{E} \sup_{1\leq a\leq b\leq n}\left\|\sum_{t=a}^{b} Z_t\right\| \leq 2\,\mathbb{E} \sup_{1\leq b\leq n}\left\|\sum_{t=1}^{b} Z_t\right\|.$$

Observe that the sequence $X_\tau := \|\sum_{t=1}^{\tau} Z_t\|$ is clearly a non-negative submartingale (with respect to (Z_t)), since

$$\mathbb{E}[X_\tau \mid Z_1, \ldots Z_{\tau-1}] = \mathop{\mathbb{E}}_{Z_\tau}\left[\left\|\sum_{t=1}^{\tau} Z_t\right\| \mid Z_1, \ldots Z_{\tau-1}\right] \geq \left\|\sum_{t=1}^{\tau-1} Z_t + \mathop{\mathbb{E}}_{Z_\tau}[Z_\tau \mid Z_1, \ldots Z_{\tau-1}]\right\| = X_{\tau-1}.$$

This, combined with boundedness of $\|Z_t\|$, means that we can apply Corollary 9 with $\eta = 1$, which gives

$$\mathbb{E} \sup_{1\leq b\leq n}\left\|\sum_{t=1}^{b} Z_t\right\| \leq \log(Bn)\,\mathbb{E}\left\|\sum_{t=1}^{n} Z_t\right\| + 1.$$

□

UMD Inequalities

Theorem 17 (Hytönen et al. (2016), Theorem 4.2.7)**.** *Suppose* $(\mathfrak{B}, \|\cdot\|)$ *is such that the deterministic UMD inequality*

$$\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t dZ_t\right\|^p \leq \mathbf{C}_p^p\,\mathbb{E}\left\|\sum_{t=1}^{n} dZ_t\right\|^p$$

holds for $p \in (1, \infty)$*. Then the determinstic UMD inequality*

$$\mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t dZ_t\right\|^q \leq \mathbf{C}_q^q\,\mathbb{E}\left\|\sum_{t=1}^{n} dZ_t\right\|^q$$

holds for any $q \in (1, \infty)$*, with*

$$\mathbf{C}_q \leq 100\left(\frac{q}{p} + \frac{q'}{p'}\right)\mathbf{C}_p.$$

Theorem 18 (Pisier (2011), Theorem 8.23)**.** *Suppose that the deterministic UMD inequality*

$$\sup_n \mathbb{E}\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\|^2 \leq \mathbf{C}_2^2 \sup_n \mathbb{E}\left\|\sum_{t=1}^n dZ_t\right\|^2$$

holds for any sign sequence. Then the L_1 UMD inequality

$$\mathbb{E}\sup_n\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\| \leq 54\mathbf{C}_2\, \mathbb{E}\sup_n\left\|\sum_{t=1}^n dZ_t\right\|$$

holds as well.

Corollary 11. If deterministic UMD inequality

$$\mathbb{E}\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\|^2 \leq \mathbf{C}_2^2\, \mathbb{E}\left\|\sum_{t=1}^n dZ_t\right\|^2$$

holds for any sign sequence, then the L_1 UMD inequality

$$\mathbb{E}\sup_n\left\|\sum_{t=1}^n \epsilon_t dZ_t\right\| \leq 108\mathbf{C}_2\, \mathbb{E}\sup_n\left\|\sum_{t=1}^n dZ_t\right\|$$

holds as well.

Theorem 19 (Hytönen et al. (2016), Proposition 4.2.17)**.** *If* $(\mathfrak{B}, \|\cdot\|)$ *is* UMD_p *with constant* $\mathbf{C}_p$*, then* $(\mathfrak{B}^\star, \|\cdot\|_\star)$ *is* $\mathrm{UMD}_{p'}$ *with constant* $\mathbf{C}_{p'} = \mathbf{C}_p$*.*

8.7.6 Burkholder/Bellman Functions

Elementary Design of Zig-Zag Concave Burkholder Functions

The following construction for the scalar case does not obtain optimal constants, but should give the reader a taste of how one can construct a zig-zag concave Burkholder function from first principles.

Theorem 20 (Elementary Scalar $\mathbf{U}$ Function)**.** *Let $k \geq 4$ be an even integer. Then the function*

$$\mathbf{U}(x,y) = \frac{k}{2}\left(x^k - 2\binom{k}{2}x^{k-2}y^2 - \frac{1}{k-2}\binom{k}{2}^{-1}\left(4\binom{k}{2}\binom{k-2}{2}\right)^{k-2}y^k\right).$$

is Zig-Zag for $|\cdot|^k$, with UMD constant

$$\mathbf{C}_k \leq \alpha k^4$$

for some constant α.

Proof. Let $\widetilde{\mathbf{U}}(x,y) = x^k - Cx^{k-2}y^2 - By^k$. We will show that $\widetilde{\mathbf{U}}$ is Zig-Zag for an appropriate choice of constants B and C.

Fix $h \in \mathbb{R}$ and let $G(t) = \widetilde{\mathbf{U}}(x+ht, y+\epsilon ht)$ for $\epsilon \in \{\pm 1\}$. By direct calculation we have

$$G''(0) = 2h^2\left[\binom{k}{2}x^{k-2} - C\left(\binom{k-2}{2}x^{k-4}y^2 + 2\binom{k-2}{2}\epsilon x^{k-3}y + x^{k-2}\right) - B\binom{k}{2}y^{k-2}\right]$$

Since k is even, $x^{k-4}y^2$ is a square; we will simply drop this term.

$$\begin{aligned}&\leq 2h^2\left[\binom{k}{2}x^{k-2} - C\left(2\binom{k-2}{2}\epsilon x^{k-3}y + x^{k-2}\right) - B\binom{k}{2}y^{k-2}\right]\\ &\leq 2h^2\left[\binom{k}{2}x^{k-2} + 2C\binom{k-2}{2}|x|^{k-3}|y| - Cx^{k-2} - B\binom{k}{2}y^{k-2}\right]\end{aligned}$$

By Young's inequality, we have

$$2C\binom{k-2}{2}|x|^{k-3}|y| = \underbrace{\left(2C\binom{k-2}{2}|y|\right)}_{a} \cdot \underbrace{|x|^{k-3}}_{b},$$

where we have applied $a \cdot b \leq \frac{1}{k-2}a^{k-2} + \frac{k-3}{k-2}b^{\frac{k-2}{k-3}}$. This expression is at most $\frac{1}{k-2}\left((2C\binom{k-2}{2})^{k-2}y^{k-2} + (k-3)x^{k-2}\right)$.

Returning to $G''(0)$, we now have

$$G''(0) \leq 2h^2\left[\left(\binom{k}{2} + \frac{k-3}{k-2} - C\right)x^{k-2} + \left(\frac{1}{k-2}\left(2C\binom{k-2}{2}\right)^{k-2} - B\binom{k}{2}\right)y^{k-2}\right].$$

In particular, we can take $C \geq 2\binom{k}{2}$ and $B \geq \frac{1}{k-2}\left(2C\binom{k-2}{2}\right)^{k-2}\binom{k}{2}^{-1}$.

$$\leq 0.$$

This certifies that G is zig-zag concave. To see the upper bound property, observe by that Young's inequality,

$$x^k - Cx^{k-2}y^2 - By^k \geq \frac{2}{k}x^k - \left(\frac{2}{k}C^{\frac{k}{2}} + B\right)y^k.$$

Hence, if we take $\mathbf{U}(x, y) = \frac{k}{2}\widetilde{\mathbf{U}}(x, y)$, we have

$$\mathbf{U}(x, y) \geq x^k - \left(C^{\frac{k}{2}} + \frac{k}{2}B\right)y^k.$$

$\square$

Zig-Zag Concave Burkholder Functions with Exponent $p = 1$

Definition 6 **($(1, 1)$ Weak Type Burkholder Function).** *A function $\mathbf{U} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ is $(\|\cdot\|, \beta)$ Zig-Zag for weak type if*

1. $\mathbf{U}(x, x') \geq \mathbb{1}\{\|x\| \geq 1\} - \beta\|x'\|$.
2. $\mathbf{U}$ *is* zig-zag concave*: $z \mapsto \mathbf{U}(x + \epsilon z, x' + z)$ is concave for all $x, x' \in \mathcal{X}$ and $\epsilon \in \{\pm 1\}$.*
3. $\mathbf{U}(0, 0) \leq 0$.

Lemma 15. Suppose we are given a weak type Burkholder function $\mathbf{U}_{\|\cdot\|,\mathrm{weak}}$ for $(\|\cdot\|, \beta)$. Then for all arguments x, y with $\|x\|, \|y\| \leq B$, the following function is Zig-Zag for $(\|\cdot\|, 1, C\beta\log(B/\epsilon))$ up to additive slack ϵ:

$$\mathbf{U}_{\|\cdot\|,1}(x, y) := \epsilon \sum_{k=1}^{N} \mathbf{U}_{\|\cdot\|,\mathrm{weak}}(x/\lambda_k, y/\lambda_k), \tag{8.36}$$

where $N = \lceil B/\epsilon \rceil$ and $\lambda_k = k\epsilon$.

Proof of Lemma 15. Let $V(x, y) = \|x\| - C'\beta \log(B/\epsilon)\|y\| - \epsilon$. We will show that $\mathbf{U}(x, y) \geq V(x, y)$ when $\|x\|, \|y\| \leq B$.

$$\begin{aligned}
V(x, y) &= \|x\| - C'\beta \log(B/\epsilon)\|y\| - \epsilon \\
&\leq \epsilon + \epsilon \sum_{k=1}^{N} \mathbb{1}\{\|x\| \geq \lambda_k\} - C'\beta \log(B/\epsilon)\|y\| - \epsilon \\
&\leq \epsilon \sum_{k=1}^{N} \left[\mathbf{U}_{\|\cdot\|,\mathrm{weak}}(x/\lambda_k, y/\lambda_k) + \frac{\beta}{\lambda_k}\|y\| \right] - C'\beta \log(B/\epsilon)\|y\| \\
&= \mathbf{U}_{\|\cdot\|,1}(x, y) + \epsilon \sum_{k=1}^{N} \frac{\beta}{\lambda_k}\|y\| - C'\beta \log(B/\epsilon)\|y\| \\
&= \mathbf{U}_{\|\cdot\|,1}(x, y) + \beta\|y\| \sum_{k=1}^{N} \frac{1}{k} - C'\beta \log(B/\epsilon)\|y\| \\
&\leq \mathbf{U}_{\|\cdot\|,1}(x, y) + C\beta\|y\| \log(N) - C'\beta \log(B/\epsilon)\|y\|
\end{aligned}$$

For sufficiently large C':

$$\leq \mathbf{U}_{\|\cdot\|,1}(x, y).$$

It can be seen immediately that $\mathbf{U}_{\|\cdot\|,1}(x, y)$ is zig-zag concave and has $\mathbf{U}_{\|\cdot\|,1}(0, 0) \leq 0$. $\square$

Zig-Zag Concavity and ζ-Convexity

Definition 7. *We say $(\mathfrak{B}, \|\cdot\|)$ is ζ-convex if there exists $\zeta : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$ such that*

1. *ζ is biconvex.*
2. *$\zeta(x, y) \leq \|x + y\| \quad$ if $\|x\| = \|y\| = 1$,*

Given a such a function ζ, we can construct a "canonical" function u which satisfies some additional properties

Definition 8.

$$u(x,y) := \begin{cases} \max\{\zeta(x,y), \|x+y\|\}, & \max\{\|x\|,\|y\|\} < 1 \\ \|x+y\|, & \max\{\|x\|,\|y\|\} \geq 1. \end{cases}.$$

Then u is biconvex, has $\zeta(0,0) \leq u(0,0)$, and satisfies

$$u(x,y) \leq \|x+y\| \quad \text{if } \max\{\|x\|,\|y\|\} \geq 1.$$

Also, $u(x,y) = u(-x,-y)$.

Assumption 2. $u(x,-x) \leq 0$.

The ζ function given in Example 20 satisfies this condition. More generally, most ζ functions can be made to satisfy this property with a slight blowup in the UMD constant they imply (c.f. (Burkholder, 1986, Lemma 8.5)).

By (Burkholder, 1986, 8.6) Assumption 2 implies $u(x,y) \leq u(0,0) + \|x+y\|$. The following argument due to (Burkholder, 1986) shows how to create a $\mathbf{U}$ function from the function u.

Theorem 21. *Suppose $\|\cdot\|$ is ζ-convex and u satisfies Assumption 2. Then this space is UMD with weak type estimate*

$$\mathbb{P}\left(\left\|\sum_{t=1}^{n} dZ_t\right\| \geq 1\right) \leq \frac{2}{u(0,0)} \mathbb{E}\left\|\sum_{t=1}^{n} \epsilon_t dZ_t\right\|$$

for any martingale difference sequence (dZ_t). Furthermore, the function

$$\mathbf{U}(x,y) = 1 - \frac{u(x+y, y-x)}{u(0,0)}$$

is weak-type Zig-Zag for $(\|\cdot\|, \frac{2}{\zeta(0,0)})$, in the sense of Definition 6.

Proof of Theorem 21. For the weak type estimate, we will start with the base function

$$V(x,y) = \mathbb{1}\{\|x\| \geq 1\} - \frac{2}{u(0,0)}\|y\|.$$

We will now show that $V(x,y) \le \mathbf{U}(x,y)$. First, observe that

$$\mathbb{1}\{\|x\| \ge 1\} = \mathbb{1}\{\|(x+y)+(x-y)\| \ge 2\} \le \mathbb{1}\{\max\{\|x+y\|, \|y-x\|\} \ge 1\}$$
$$\le \mathbb{1}\{2\|y\| \ge u(x+y, y-x)\},$$

where the last inequality follows from the additional property of u from Definition 8. We have now established

$$\begin{aligned} V(x,y) &\le \mathbb{1}\{2\|y\| \ge u(x+y,y-x)\} - \frac{2}{u(0,0)}\|y\| \\ &= \mathbb{1}\{2\|y\| - u(x+y,y-x) + u(0,0) \ge u(0,0)\} - \frac{2}{u(0,0)}\|y\| \end{aligned}$$

By the second additional property of u from Definition 8, $2\|y\| - u(x+y, y-x) + u(0,0) \ge 0$, and so we may apply Markov's inequality

$$\begin{aligned} &\le \frac{2\|y\| - u(x+y,y-x) + u(0,0)}{u(0,0)} - \frac{2}{u(0,0)}\|y\| \\ &= \mathbf{U}(x,y). \end{aligned}$$

Observe that $\mathbf{U}(0,0) = 0$ and, since u is biconvex, $-u(x+y, y-x)$ is zig-zag concave, and so $\mathbf{U}$ is itself zig-zag concave. We can now prove that the UMD property holds with constant $\frac{2}{u(0,0)} \le \frac{2}{\zeta(0,0)}$ using the standard step-by-step peeling argument with $\mathbf{U}$ described in Hytönen et al. (2016), Theorem 4.5.6. □

Example 20 (ℓ_1^d Osekowski (2016))**.** *Define*

$$z(x,y) = \begin{cases} \frac{a\langle x,y\rangle}{2} - \frac{1}{2a}, & \|x+y\| + \|x-y\| \le 2/a \\ \frac{\|x+y\|}{2}\log\left(\frac{a}{2}(\|x+y\|+\|x-y\|)\right) - \frac{\|x-y\|}{2}, & \|x+y\| + \|x-y\| > 2/a \end{cases}.$$

Then define

$$\zeta(x,y) = \frac{2}{\log(3a)}\left(1 + \sum_{i=1}^{d} z(x_i, y_i)\right).$$

For $a \ge d\log d$ the ζ-convexity properties are satisfied and the bound

$$\zeta(0,0) \le \frac{2}{\log d + \log(2\log d)}\left(1 - \frac{1}{2\log d}\right)$$

is achieved.

8.8 Chapter Notes

This chapter is based on Foster et al. (2017b). A number of questions centered around the zig-zag concave Burkholder functions remain open. Here we state two.

General function classes Understanding what abstract properties of the function class $\mathcal{F}$ (behavior of covering numbers etc.) lead to the generalized UMD inequality is an important problem. Progress in this direction will hopefully lead to more concrete examples of classes with the property.

Tighter rates for specific losses The empirical Rademacher complexity regret bound is not tight for strongly convex losses such as the square loss. *Offset rademacher complexity* techniques have been used to obtain tight worst-case rates in this case (Rakhlin and Sridharan, 2014). Developing UMD-type inequalities for the offset Rademacher complexity would yield new adaptive algorithms for a number of settings.

CHAPTER 9
ONLINE OPTIMIZATION

In this chapter we turn our focus to online convex optimization. The online convex optimization model has seen widespread use for solving large-scale empirical risk minimization problems for machine learning (Zinkevich, 2003; Duchi et al., 2011). Beyond empirical risk minimization, online convex optimization algorithms give way to stochastic convex optimization guarantees, and in imply upper bounds on the oracle complexity of stochastic optimization (Nemirovski et al., 1983).

We introduce a very general type of adaptivity for online convex optimization inspired by *model selection*-based adaptivity results in statistical learning. We call this notion of adaptivity *online model selection*, as it subsumes the classical model selection framework and extends it to the online convex optimization setting. The new notion of adaptivity also encompasses literature on *parameter-free* online optimization (McMahan and Abernethy, 2013; McMahan and Orabona, 2014; Orabona, 2014; Orabona and Pál, 2016) which focuses on adapting to a single scalar parameter for optimization in Hilbert space, but is substantially more general.

The results in this chapter leverage the equivalence of adaptive online learning and martingale inequalities to a) characterize the achievability of online model selection adaptivity via a type of martingale inequality we call a *multi-scale maximal inequality* and b) derive a new efficient online convex optimization algorithm that achieves this form of adaptivity. The core algorithmic tool is a new *multi-scale* algorithm for prediction with expert advice based on random playout. This can be seen as an algorithmic realization of the achievability result presented in Section 6.4.1. Applications include new online model selection guarantees for matrix classes, non-nested convex sets, and $\mathbb{R}^d$ with generic regularizers.

Finally, we generalize these results by providing oracle inequalities for arbitrary non-linear classes in the online supervised learning model.

9.1 Background

A key problem in the design of learning algorithms is the choice of the hypothesis set $\mathcal{F}$. This is known as the *model selection* problem. The choice of $\mathcal{F}$ is driven by inherent trade-offs. In the statistical learning setting, this can be analyzed in terms of the *estimation* and *approximation errors*. A richer or more complex $\mathcal{F}$ helps better approximate the Bayes predictor (smaller approximation error). On the other hand, a hypothesis set that is too complex may have too large a VC dimension or have unfavorable Rademacher complexity, thereby resulting in looser guarantees on the difference between the loss of a hypothesis and that of the best in class (large estimation error).

In the batch setting, this problem has been extensively studied with the main ideas originating in the seminal work of Vapnik and Chervonenkis (1971) and Vapnik (1982) and the principle of Structural Risk Minimization (SRM). This is typically formulated as follows: let $(\mathcal{F}_i)_{i\in\mathbb{N}}$ be an infinite sequence of hypothesis sets (or models); the problem consists of using the training sample to select a hypothesis set $\mathcal{F}_i$ with a favorable estimation-approximation trade-off and choosing the best hypothesis f in $\mathcal{F}_i$.

If we had access to a hypothetical oracle informing us of the best choice of i for a given instance, the problem would reduce to the standard one of learning with a fixed hypothesis set. Remarkably though, techniques such as SRM or similar penalty-based model selection methods return a hypothesis f^* that enjoys finite-sample learning guarantees that are almost as favorable as those that would be obtained had an oracle informed us of the index i^* of the best-in-class classifier's hypothesis set (Vapnik, 1982; Devroye et al., 1996;

Shawe-Taylor et al., 1998; Koltchinskii, 2001; Bartlett et al., 2002; Massart, 2007). Such guarantees are sometimes referred to as *oracle inequalities*. They can be derived even for data-dependent penalties (Koltchinskii, 2001; Bartlett et al., 2002; Bartlett and Mendelson, 2003).

A line in of research in online optimization community suggests that similar results might be possible in online optimization (and thus stochastic optimization as well). Specifically, McMahan and Abernethy (2013); McMahan and Orabona (2014); Orabona (2014); Orabona and Pál (2016) all present algorithms that efficiently achieve model selection oracle inequalities for the important special case where $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is a sequence of nested balls in a Hilbert space. Such results naturally raise the following questions for the online setting: can we develop a *general* theory of model selection in online convex optimization that works for arbitrary model sequences, not just Hilbert spaces? Moreover, can we achieve such guarantees *efficiently*? Note that unlike the statistical setting, in online learning one cannot split samples to first learn the optimal predictor within each subclass and then later learn the optimal subclass choice, so new algorithmic ideas are required.

The first approach that a familiar reader might think of for tackling the online model selection problem is to run for each i an online learning algorithm that minimizes regret against $\mathcal{F}_i$, and then aggregate over these algorithms using the multiplicative weights algorithm for prediction with expert advice. This would work if all the losses or "experts" considered were uniformly bounded by a reasonably small quantity. However, for problems in online convex optimization the losses of predictors or experts for each $\mathcal{F}_i$ may grow with i. Using simple aggregation would scale our regret with the magnitude of the largest $\mathcal{F}_i$ and not the i^* we want to compare against. This is the main technical challenge faced in this context, and one that we fully address in this paper.

Our results are based on a novel *multi-scale algorithm* for prediction with expert advice. This algorithm works in a situation where the different experts' losses lie in different ranges, and guarantees that the regret to each individual expert is adapted to the range of its losses. The algorithm can also take advantage of a given prior over the experts reflecting their importance. This general, abstract setting of prediction with expert advice yields generic online model selection algorithms. The result is achieved by exploiting the equivalence developed in Part II: We first characterize the achievability of the multi-scale prediction guarantee by proving a certain "multi-scale maximal inequality" for martingales, and then use minimax analysis to derive an efficient algorithm that achieves this form of adaptivity.

9.1.1 Preliminaries

Setup and Goals. We work in the online convex optimization setting (Protocol 3), where the learner selects decisions from a convex subset $\mathcal{W}$ of some Banach space $\mathfrak{B}$. Regret to a comparator $w \in \mathcal{W}$ in this setting is defined as $\mathrm{Reg}_n(w) = \sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w)$.

Suppose $\mathcal{W}$ can be decomposed into sets $\mathcal{W}_1, \mathcal{W}_2, \ldots$. For a fixed set $\mathcal{W}_k$, the optimal regret, if one tailors the algorithm to compete with $\mathcal{W}_k$, is typically characterized by some measure of intrinsic complexity of the class such as Littlestone's dimension or sequential Rademacher complexity (Ben-David et al., 2009; Rakhlin et al., 2014), denoted $\mathbf{Comp}_n(\mathcal{W}_k)$. We would like to develop adaptive algorithms for online convex optimization that produce a sequence $(w_t)_{t\geq 1}$ such that

$$\sum_{t=1}^{n} f_t(w_t) - \min_{w\in\mathcal{W}_k} \sum_{t=1}^{n} f_t(w) \leq \mathbf{Comp}_n(\mathcal{W}_k) + \mathbf{Pen}_n(k) \quad \forall k. \tag{9.1}$$

This equation is called an *oracle inequality* and states that the performance of the sequence (w_t) matches that of a comparator that minimizes the bias-variance tradeoff $\min_k\{\min_{w\in\mathcal{W}_k} \sum_{t=1}^{n} f_t(w) + \mathbf{Comp}_n(\mathcal{W}_k)\}$, up to a penalty $\mathbf{Pen}_n(k)$ whose scale ideally matches that of $\mathbf{Comp}_n(\mathcal{W}_k)$. We shall see shortly that ensuring that the scale of $\mathbf{Pen}_n(k)$

does indeed match is the core technical challenge in developing online oracle inequalities for commonly used classes.

9.2 Online Model Selection

9.2.1 Multi-Scale Aggregation

Let us briefly motivate the main technical challenge overcome by the model selection approach we consider. The most widely studied oracle inequality in online learning has the following form

$$\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w) \leq O\Big((\|w\|_2 + 1)\sqrt{n \cdot \log((\|w\|_2 + 1)n)}\Big) \quad \forall w \in \mathbb{R}^d. \tag{9.2}$$

In light of (9.1), a *model selection* approach to obtaining this inequality would be to split the set $\mathcal{W} = \mathbb{R}^d$ into ℓ_2 norm balls of doubling radius, i.e. $\mathcal{W}_k = \left\{w \mid \|w\|_2 \leq 2^k\right\}$. A standard fact (Hazan, 2016) is that such a set has $\mathbf{Comp}_n(\mathcal{W}_k) = 2^k\sqrt{n}$ if one optimizes over it using Mirror Descent, and so obtaining the oracle inequality (9.1) is sufficient to recover (9.2), so long as $\mathbf{Pen}_n(k)$ is not too large relative to $\mathbf{Comp}_n(\mathcal{W}_k)$.

We view online model selection as a problem of prediction with expert advice (Cesa-Bianchi and Lugosi, 2006), where the experts correspond to the different model classes one is choosing from. Our basic meta-algorithm, MULTISCALEFTPL (Algorithm 6), operates in the following setup. The algorithm has access to a finite number, N, of experts. In each round, the algorithm is required to choose one of the N experts. Then the losses of all experts are revealed, and the algorithm incurs the loss of the chosen expert.

The twist from the standard setup is that the losses of all the experts are *not* uniformly bounded in the same range. Indeed, for the setup described for the oracle inequality (9.2), class $\mathcal{W}_k$ will produce predictions with norm as large as 2^k. Therefore, here, we assume that expert i incurs losses in the range $[-c_i, c_i]$, for some known parameter $c_i \geq 0$. The goal is to design an online learning algorithm whose regret to expert i scales with c_i, rather than $\max_i c_i$, which is what out-of-the-box algorithms for learning from expert advice (such as the multiplicative weights strategy or AdaHedge (De Rooij et al., 2014)) would achieve. Indeed, any regret bound scaling in $\max_i c_i$ will be far too large to achieve (9.2), as the term $\mathbf{Pen}_n(k)$ will dominate. This new type of scale-sensitive regret bound, which is achieved by our algorithm MULTISCALEFTPL, is stated below.

Algorithm 6

procedure MULTISCALEFTPL(c, π) ▷ Scale vector c with $c_i \geq 1$, prior distribution π.
  for time $t = 1, \dots, n$: **do**
    Draw sign vectors $\sigma_{t+1}, \dots, \sigma_n \in \{\pm 1\}^N$ each uniformly at random.
    Compute distribution

$$p_t(\sigma_{t+1:n}) = \arg\min_{p \in \Delta_N} \sup_{g_t : |g_t[i]| \leq c_i} \left[\langle p, g_t \rangle + \sup_{i \in [N]} \left[-\sum_{s=1}^{t} \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^{n} \sigma_s[i] c_i - \mathcal{B}(i) \right] \right],$$

    where $\mathcal{B}(i) = 5 c_i \sqrt{n \log(4 c_i^2 n / \pi_i)}$.
    Play $i_t \sim p_t$.
    Observe loss vector g_t.
  end for
end procedure

Theorem 22. *Suppose the loss sequence $(g_t)_{t \leq n}$ satisfies $|g_t[i]| \leq c_i$ for a sequence $(c_i)_{i \in [N]}$ with each $c_i \geq 1$. Let $\pi \in \Delta_N$ be a given prior distribution on the experts. Then, playing the strategy $(p_t)_{t \leq n}$ given by Algorithm 6,* MULTISCALEFTPL *yields the following regret bound:*[1]

$$\mathbb{E}\left[\sum_{t=1}^{n} \langle e_{i_t}, g_t \rangle - \sum_{t=1}^{n} \langle e_i, g_t \rangle \right] \leq O\Big(c_i \sqrt{n \log(n c_i / \pi_i)} \Big) \quad \forall i \in [N]. \tag{9.3}$$

[1]This regret bound holds under expectation over the player's randomization. It is assumed that each g_t is selected before the randomized strategy p_t is revealed, but may adapt to the distribution over p_t. In fact, a slightly stronger version of this bound holds, namely $\mathbb{E}\Big[\sum_{t=1}^{n} \langle e_{i_t}, g_t \rangle - \min_{i \in [N]} \Big\{ \sum_{t=1}^{n} \langle e_i, g_t \rangle + O\Big(c_i \sqrt{n \log(n c_i / \pi_i)} \Big) \Big\} \Big] \leq 0$. A similar strengthening applies to all subsequent bounds.

Briefly, the key to showing achievability of this theorem is the following maximal inequality, which arises via the equivalence of adaptive prediction inequalities and martingale inequalities derived in Part II. The inequality is multi-scale analogue of the classical maximal inequality subgaussian random variables (e.g. Boucheron et al. (2013)).

Lemma 16 (Multi-Scale Maximal Inequality)**.** Let $(X_i)_{i\in[N]}$ be a real-valued random process for which there exists a sequence $(h_i)_{i\in[N]}$ with $h_i > 0$ such that the moment generating function bound $\mathbb{E}\, e^{\lambda X_i} \leq e^{\lambda^p h_i}$ is satisfied for all $\lambda > 0$ and some choice of $p > 0$. Then for any distribution $\pi \in \Delta_N$ for which $h_i/\pi_i \geq e$ for all $i \in [N]$ it holds that

$$\mathbb{E} \max_{i\in[N]} \Big\{ X_i - (2 + 1/p) h_i^{1/p} (\log(h_i) + \log(1/\pi_i))^{1-1/p} \Big\} \leq \sum_{i\in[N]} \frac{\pi_i}{h_i}. \tag{9.4}$$

For our application the inequality arises when each X_i is a martingale difference sequence with increments bounded in magnitude by c_i, and the key takeaway is that if we look at the maximum deviation relative to an offset function $\mathcal{B}(i)$, the right-hand side need not scale with $\max_i c_i$.

Compared to related FTPL algorithms (Rakhlin et al., 2012), the analysis of Theorem 22 is surprisingly delicate, as additive c_i factors at any point in the analysis can spoil the desired regret bound (9.3) if different c_is differ by orders of magnitude.

The min-max optimization problem in MULTISCALEFTPL can be solved in $\widetilde{O}(N^{3.5})$ time using linear programming. See Section 9.3.1 for a detailed discussion.

9.2.2 Adaptive Algorithms for Online Convex Optimization

One can readily apply MULTISCALEFTPL for online optimization problems whenever it is possible to bound the losses of the different experts a-priori. One such application is to online convex optimization, where each "expert" is a a particular OCO algorithm, and for which such a bound can be obtained via appropriate bounds on the relevant norms of the parameter vectors and the gradients of the loss functions. We detail this application — which yields algorithms for parameter-free online learning and more — below. All of the algorithms in this section are derived using a unified meta-algorithm strategy MULTISCALEOCO.

The setup is as follows. We have access to N sub-algorithms, denoted $\textsc{Alg}_i$ for $i \in [N]$. In round t, each sub-algorithm $\textsc{Alg}_i$ produces a prediction $w_t^i \in \mathcal{W}_i$, where $\mathcal{W}_i$ is a set in a vector space V over $\mathbb{R}$ containing 0. Our meta-algorithm is then required to choose one of the predictions w_t^i. Then, a loss function $f_t : V \to \mathbb{R}$ is revealed, whereupon $\textsc{Alg}_i$ incurs loss $f_t(w_t^i)$, and the meta-algorithm suffers the loss of the chosen prediction. We make the following assumption on the sub-algorithms:

Assumption 3. The sub-algorithms satisfy the following conditions:

- For each $i \in [N]$, there is an associated norm $\|\cdot\|_{(i)}$ such that $\sup_{w \in \mathcal{W}_i} \|w\|_{(i)} \leq R_i$.
- For each $i \in [N]$, the sequence of functions f_t are L_i-Lipschitz on $\mathcal{W}_i$ with respect to $\|\cdot\|_{(i)}$.
- For each sub-algorithm $\textsc{Alg}_i$, the iterates $(w_t^i)_{t \leq n}$ enjoy a regret bound $\sum_{t=1}^n f_t(w_t^i) - \inf_{w \in \mathcal{W}_i} \sum_{t=1}^n f_t(w) \leq \mathrm{Reg}_n(i)$, where $\mathrm{Reg}_n(i)$ may be data- or algorithm-dependent.

In most applications, $\mathcal{W}_i$ will be a convex set and f_t a convex function; this convexity is not necessary to prove a regret bound for the meta-algorithm. We simply need boundedness of the set $\mathcal{W}_i$ and Lipschitzness of the functions f_t, as specified in Assumption 3. This assumption implies that for any i, we have $|f_t(w) - f_t(0)| \leq R_i L_i$ for any $w \in \mathcal{W}_i$. Thus, we can design a meta-algorithm for this setup by using MULTISCALEFTPL with $c_i = R_i L_i$, which is precisely what is described in Algorithm 7. The following theorem provides a bound on the regret of MULTISCALEOCO; a direct consequence of Theorem 22.

Algorithm 7

```
procedure MultiScaleOCO({Alg_i, R_i, L_i}_{i∈[N]}, π)    ▷ Collection of sub-algorithms, prior π.
    c ← (R_i · L_i)_{i∈[N]}                              ▷ Sub-algorithm scale parameters.
    for t = 1, ..., n do
        w_t^i ← Alg_i(f̃_1, ..., f̃_{t-1}) for each i ∈ [N].
        i_t ← MultiScaleFTPL[c, π](g_1, ..., g_{t-1}).
        Play w_t = w_t^{i_t}.
        Observe loss function f_t and let f̃_t(w) = f_t(w) − f_t(0).
        g_t ← (f̃_t(w_t^i))_{i∈[N]}.
    end for
end procedure
```

Theorem 23. *Without loss of generality, assume that $R_iL_i \geq 1$*[2]*. Suppose that the inputs to Algorithm 7 satisfy Assumption 3. Then the iterates $(w_t)_{t\leq n}$ returned by Algorithm 7 follow the regret bound*

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t) - \inf_{w\in\mathcal{W}_i}\sum_{t=1}^{n} f_t(w)\right] \leq \mathbb{E}[\mathrm{Reg}_n(i)] + O\left(R_iL_i\sqrt{n\log(R_iL_in/\pi_i)}\right) \quad \forall i \in [N]. \tag{9.5}$$

Theorem 23 shows that if we use Algorithm 7 to aggregate the iterates produced by a collection of sub-algorithms $(\textsc{Alg}_i)_{i\in[N]}$, the regret against any sub-algorithm i will only depend on that algorithm's scale, not the regret of the worst sub-algorithm.

Application 1: Parameter-free Online Learning in Uniformly Convex Banach Spaces. As the first application of our framework, we give a generalization of the parameter-free online learning bounds found in McMahan and Abernethy (2013); McMahan and Orabona (2014); Orabona (2014); Orabona and Pál (2016); Cutkosky and Boahen (2016) from Hilbert spaces to arbitrary uniformly convex Banach spaces. Recall that a Banach space $(\mathfrak{B}, \|\cdot\|)$ is $(2,\lambda)$-uniformly convex if $\frac{1}{2}\|\cdot\|^2$ is λ-strongly convex with respect to itself (Pisier, 2011). Our algorithm obtains a generalization of the oracle inequality (9.2) for any uniformly convex

[2]For notational convenience all Lipschitz bounds are assumed to be at least 1 without loss of generality for the remainder of the chapter.

$(\mathfrak{B}, \|\cdot\|)$ by running multiple instances of *Mirror Descent* — the workhorse of online convex optimization — and aggregating their iterates using MULTISCALEOCO. This strategy is thus efficient whenever Mirror Descent can be implemented efficiently. The collection of sub-algorithms used by MULTISCALEOCO, which was alluded to at the beginning of this section is as follows: For each $1 \leq i \leq N := n+1$, set $R_i = e^{i-1}$, $L_i = L$, $\mathcal{W}_i = \{w \in \mathfrak{B} \mid \|w\| \leq R_i\}$, $\eta_i = \frac{R_i}{L}\sqrt{\frac{\lambda}{n}}$, and $\textsc{Alg}_i = \textsc{MirrorDescent}(\eta_i, \mathcal{W}_i, \|\cdot\|^2)$. Finally, set $\pi = \mathrm{Uniform}([n+1])$.

Mirror Descent is reviewed in detail in Section 9.3.2, but the only feature of its performance of importance to our analysis is that, when configured as described above, the iterates $(w_t^i)_{t\leq n}$ produced by $\textsc{Alg}_i$ specified above will satisfy $\sum_{t=1}^{n} f_t(w_t^i) - \inf_{w\in\mathcal{W}_i}\sum_{t=1}^{n} f_t(w) \leq O(R_i L\sqrt{\lambda n})$ on any sequence of losses that are L-Lipschitz with respect to $\|\cdot\|_\star$. Using just this simple fact, combined with the regret bound for MULTISCALEOCO and a few technical calculations, we can deduce the following parameter-free learning oracle inequality:

Theorem 24 (Oracle inequality for uniformly convex Banach spaces)**.** *The iterates $(w_t)_{t\leq n}$ produced by* MULTISCALEOCO *on any L-Lipschitz (w.r.t. $\|\cdot\|_\star$) sequence of losses $(f_t)_{t\leq n}$ satisfy*

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w)\right] \leq O\Big(L\cdot(\|w\|+1)\sqrt{n\cdot\,\log((\|w\|+1)Ln)/\lambda}\Big) \quad \forall w \in \mathfrak{B}. \quad (9.6)$$

Note that the above oracle inequality applies for *any uniformly convex norm* $\|\cdot\|$. Previous results only obtain bounds of this form efficiently when $\|\cdot\|$ is a Hilbert space norm or ℓ_1. As is standard for such oracle inequality results, the bound is weaker than the optimal bound if $\|w\|$ were selected in advance, but only by a mild $\sqrt{\log((\|w\|+1)Ln)}$ factor.

Proposition 14. The algorithm can be implemented in time $O(T_{\mathrm{MD}}\cdot\mathrm{poly}(n))$ per iteration, where T_{MD} is the time complexity of a single Mirror Descent update.

In the example above, the $(2,\lambda)$-uniform convexity condition was mainly chosen for familiarity. The result can easily be generalized to related notions such as q-uniform convexity (see Srebro et al. (2011)). More generally, the approach can be used to derive oracle inequalities with respect to general strongly convex regularizer $\mathcal{R}$ defined over the space $\mathcal{W}$. Such a bound would have the form $O\Big(L \cdot \sqrt{n(\mathcal{R}(w)+1)\cdot\,\log((\mathcal{R}(w)+1)n)}\Big)$ for typical choices of $\mathcal{R}$.

This example captures well-known *quantile bounds* (Koolen and van Erven, 2015) when one takes $\mathcal{R}$ to be the KL-divergence and $\mathcal{W}$ to be the simplex, or, in the matrix case, takes $\mathcal{R}$ to be the quantum relative entropy and $\mathcal{W}$ to be the set of density matrices, as in Hazan et al. (2012).

Application 2: Oracle Inequality for Many ℓ_p Norms. It is instructive to think of MULTISCALEOCO as executing a (scale-sensitive) online analogue of the structural risk minimization principle. We simply specify a set of subclasses and a prior π specifying the importance of each subclass, and we are guaranteed that the algorithm's performance matches that of each sub-class, plus a penalty depending on the prior weight placed on that subclass. The advantage of this approach is that the nested structure used in the Theorem 24 is completely inessential. This leads to the exciting prospect of developing parameter-free algorithms over new and exotic set systems. One such example is given now: The MULTISCALEOCO framework allows us to obtain an oracle inequality with respect to *many ℓ_p norms in $\mathbb{R}^d$ simultaneously*. To the best of our knowledge all previous works on parameter-free online learning have only provided oracle inequalities for a single norm.

Theorem 25. *Fix $\delta > 0$. Suppose that the loss functions $(f_t)_{t\leq n}$ are L_p-Lipschitz w.r.t. $\|\cdot\|_{p'}$ for each $p \in [1+\delta, 2]$, where p' is such that $\frac{1}{p}+\frac{1}{p'}=1$. Then there is a computationally efficient* *algorithm that guarantees regret bound simultaneously for all $\forall w \in \mathbb{R}^d$ and for all $p \in [1+\delta, 2]$:*

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w)\right] \leq O\left((\|w\|_p + 1)L_p\sqrt{n\log((\|w\|_p + 1)L_p \log(d)n)/(p-1)}\right). \tag{9.7}$$

The configuration in the above theorem is described in full in Section 9.3.2 in the supplementary material. This strategy can be trivially extended to handle p in the range $(2, \infty)$. The inequality holds for $p \geq 1+\delta$ rather than for $p \geq 1$ because the ℓ_1 norm is not uniformly convex, but this is easily rectified by changing the regularizer at $p = 1$; we omit this for simplicity of presentation.

We emphasize that the choice of ℓ_p norms for the result above was somewhat arbitrary — any finite collection of norms will also work. For example, the strategy can also be applied to matrix optimization over $\mathbb{R}^{d\times d}$ by replacing the ℓ_p norm with the Schatten S_p norm. The Schatten S_p norm has strong convexity parameter on the order of $p-1$ (which matches the ℓ_p norm up to absolute constants (Ball et al., 1994)) so the only change to practical change to the setup in Theorem 25 will be the running time T_{MD}. Likewise, the approach applies to (p, q)-group norms as used in multi-task learning (Kakade et al., 2012).

Application 3: Adapting to Rank for Online PCA For the online PCA task, the learner predicts from a class $\mathcal{W}_k = \left\{W \in \mathbb{R}^{d\times d} \mid W \succeq 0, \|W\|_\sigma \leq 1, \langle W, I\rangle = k\right\}$. For a fixed value of k, such a class is a convex relaxation of the set of all rank k projection matrices. After producing a prediction W_t, we experience affine loss functions $f_t(W_t) = \langle I - W_t, Y_t\rangle$, where $Y_t \in \mathcal{Y} := \left\{Y \in \mathbb{R}^{d\times d} \mid Y \succeq 0, \|Y\|_\sigma \leq 1\right\}$.

We leverage an analysis of online PCA due to (Nie et al., 2013) together with MULTISCALE-OCO to derive an algorithm that competes with many values of the rank simultaneously. This gives the following result:

Theorem 26. *There is an efficient algorithm for Online PCA with regret bound*

$$\mathbb{E}\left[\sum_{t=1}^{n}\langle I - W_t, Y_t\rangle - \min_{\substack{W \text{ projection}\\ \operatorname{rank}(W)=k}} \sum_{t=1}^{n}\langle I - W, Y_t\rangle\right] \leq \widetilde{O}\big(k\sqrt{n}\big) \quad \forall k \in [d/2].$$

For a fixed value of k, the above bound is already optimal up to log factors, but it holds for all k simultaneously.

Application 4: Adapting to Norm for Matrix Multiplicative Weights In the MATRIX MULTIPLICATIVE WEIGHTS setting (Arora et al., 2012) we consider hypothesis classes of the form $\mathcal{W}_r = \left\{W \in \mathbb{R}^{d\times d} \mid W \succeq 0, \|W\|_{\Sigma} \leq r\right\}$. Losses are given by $f_t(W) = \langle W, Y_t\rangle$, where $\|Y_t\|_{\sigma} \leq 1$. For a fixed value of r, the well-known MATRIX MULTIPLICATIVE WEIGHTS strategy has regret against $\mathcal{W}_r$ bounded by $O(r\sqrt{n\log d})$. Using this strategy for fixed r as a sub-algorithm for MULTISCALEOCO, we achieve the following oracle inequality efficiently:

Theorem 27. *There is an efficient matrix prediction strategy with regret bound*

$$\mathbb{E}\left[\sum_{t=1}^{n}\langle W_t, Y_t\rangle - \sum_{t=1}^{n}\langle W, Y_t\rangle\right] \leq (\|W\|_{\Sigma} + 1)\sqrt{n\log d\log((\|W\|_{\Sigma} + 1)n))} \quad \forall W \succeq 0. \tag{9.8}$$

A Remark on Efficiency All of our algorithms that provide bounds of the form (9.6) instantiate $O(n)$ experts with MULTISCALEFTPL because, in general, the worst case w for achieving (9.6) can have norm as large as e^n. If one has an a priori bound — say B — on the range at which each f_t attains its minimum, then the number of experts be reduced to $O(\log(B))$.

9.2.3 Back to Supervised Learning

In this final section, we show that the general online optimization algorithm developed in this chapter, MultiScaleFTPL, can also be used to develop adaptive algorithms for the general online supervised learning setting (Section 2.3, Protocol 2). In the supervised learning setting we measure regret against a benchmark class $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ of functions $f : \mathcal{X} \to \mathbb{R}$. In this case, the analogue of the online convex optimization oracle inequality (9.5) has the form

$$\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}_k} \sum_{t=1}^{n} \ell(f(x_t), y_t) \leq \mathbf{Comp}_n(\mathcal{F}_k) + \mathbf{Pen}_n(k) \quad \forall k. \tag{9.9}$$

Working in this setting makes clear a key feature of the meta-algorithm approach we have developed: We can efficiently obtain online oracle inequalities for arbitrary nonlinear function classes so long as we have an efficient algorithm for each $\mathcal{F}_k$.

We obtain a supervised learning meta-algorithm by simply feeding the observed losses $\ell(\cdot, y_t)$ (which may even be non-convex) to the meta-algorithm MULTISCALEFTPL in the same fashion as MULTISCALEOCO.

The resulting strategy is called MULTISCALELEARNING. We make the following assumptions analogous to Assumption 3, which lead to the performance guarantee for MULTISCALELEARNING given in Theorem 28 below.

Assumption 4. The sub-algorithms used by MULTISCALELEARNING satisfy the following conditions:

- For each $i \in [N]$, the iterates $(\widehat{y}_t^i)_{t \leq n}$ produced by sub-algorithm ALG$_i$ satisfy $|\widehat{y}_t^i| \leq R_i$.
- For each $i \in [N]$, the function $\ell(\cdot, y_t)$ is L_i-Lipschitz on $[-R_i, R_i]$.

- For each sub-algorithm $\textsc{Alg}_i$, the iterates $(\widehat{y}_t^i)_{t\leq n}$ enjoy a regret bound $\sum_{t=1}^n \ell(\widehat{y}_t^i, y_t) - \inf_{f\in\mathcal{F}_i}\sum_{t=1}^n \ell(f(x_t), y_t) \leq \mathrm{Reg}_n(i)$, where $\mathrm{Reg}_n(i)$ may be data- or algorithm-dependent.

Theorem 28. *Suppose that the inputs to Algorithm 8 satisfy Assumption 4. Then the iterates $(\widehat{y}_t)_{t\leq n}$ produced by the algorithm enjoy the regret bound*

$$\mathbb{E}\left[\sum_{t=1}^n \ell(\widehat{y}_t^i, y_t) - \inf_{f\in\mathcal{F}_i}\sum_{t=1}^n \ell(f(x_t), y_t)\right] \leq \mathbb{E}[\mathrm{Reg}_n(i)] + O\Big(R_iL_i\sqrt{n\log(R_iL_in/\pi_i)}\Big) \quad \forall i\in[N]. \tag{9.10}$$

Algorithm 8

```
procedure MultiScaleLearning({Alg_i, R_i, L_i}_{i∈[N]}, π)  ▷ Collection of sub-algorithms, prior π.
    c ← (R_i · L_i)_{i∈[N]}                                  ▷ Sub-algorithm scale parameters.
    Define ℓ̃(ŷ, y) = ℓ(ŷ, y) − ℓ(0, y).                       ▷ Center the loss function.
    for t = 1, ..., n do
        Receive context x_t
        ŷ_t^i ← Alg_i((x_1, y_1), ..., (x_{t−1}, y_{t−1}), x_t) for each i ∈ [N].
        i_t ← MultiScaleFTPL[c, π](g_1, ..., g_{t−1}).
        Play ŷ_t = ŷ_t^{i_t}.
        Observe y_t and let g_t = (ℓ̃_t(ŷ_t^i, y_t))_{i∈[N]}.
    end for
end procedure
```

Online Penalized Risk Minimization In the statistical learning setting, oracle inequalities for arbitrary sequences of hypothesis classes $\mathcal{F}_1,\ldots,\mathcal{F}_N$ are readily available. Such inequalities are typically stated in terms of complexity parameters for the classes $(\mathcal{F}_k)$ such as VC dimension or Rademacher complexity. For the online learning setting, it is well-known that *sequential Rademacher complexity* $\mathcal{R}^{\mathrm{seq}}(\mathcal{F})$ (cf. Chapter 6) provides a sequential counterpart to these complexity measures, meaning that it generically characterizes the minimax optimal regret for Lipschitz losses. We will obtain an oracle inequality in terms of this parameter.

Assumption 5. The sequence of hypothesis classes $\mathcal{F}_1,\ldots,\mathcal{F}_N$ are such that

1. There is an efficient algorithm $\textsc{Alg}_k$ producing iterates $(\widehat{y}_t^k)_{t\leq n}$ satisfying $\sum_{t=1}^n \ell(\widehat{y}_t^k, y_t) - \inf_{f\in\mathcal{F}_k} \sum_{t=1}^n \ell(f(x_t), y_t) \leq C \cdot L \cdot \mathcal{R}^{\mathrm{seq}}(\mathcal{F}_k)$ for any L-Lipschitz loss, where C is some constant. (an algorithm with this regret is always guaranteed to exist, but may not be efficient).

2. Each $\mathcal{F}_k$ has output range $[-R_k, R_k]$, where $R_k \geq 1$ without loss of generality.

3. $\mathcal{R}^{\mathrm{seq}}(\mathcal{F}_k) = \Omega(R_k\sqrt{n})$ — this is obtained by most non-trivial classes.

Theorem 29 (Online penalized risk minimization)**.** *Under Assumption 5 there is an efficient (in N) algorithm that achieves the following regret bound for any L-Lipschitz loss:*

$$\mathbb{E}\left[\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \inf_{f\in\mathcal{F}_k} \sum_{t=1}^n \ell(f(x_t), y_t)\right] \leq O\Big(L \cdot \mathcal{R}^{\mathrm{seq}}(\mathcal{F}_k) \cdot \sqrt{\log(L \cdot \mathcal{R}^{\mathrm{seq}}(\mathcal{F}_k) \cdot k)}\Big) \quad \forall k \in [N]. \tag{9.11}$$

As in the previous section, one can derive tighter regret bounds and more efficient (e.g. sublinear in N) algorithms if $\mathcal{F}_1, \mathcal{F}_2, \ldots$ are nested.

Application: Multiple kernel learning

Theorem 30. *Let $\mathcal{H}_1, \ldots, \mathcal{H}_N$ be reproducing kernel Hilbert spaces for which each $\mathcal{H}_k$ has a kernel $\mathbf{K}$ such that $\sup_{x\in\mathcal{X}} \sqrt{\mathbf{K}(x,x)} \leq B_k$. Then there is an efficient learning algorithm that guarantees*

$$\mathbb{E}\left[\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \sum_{t=1}^n \ell(f(x_t), y_t)\right] \leq O\Big(LB_k(\|f\|_{\mathcal{H}_k} + 1)\sqrt{\log(LB_k kn(\|f\|_{\mathcal{H}_k} + 1))}\Big) \quad \forall k, \forall f \in \mathcal{H}_k$$

for any L-Lipschitz loss, whenever an efficient algorithm is available for the norm ball in each $\mathcal{H}_k$.

9.3 Detailed Proofs

9.3.1 Multi-scale FTPL algorithm

Proof of Theorem 22. Recall that $B(i) = 5c_i\sqrt{n(\log(1/\pi_i) + \log(4c_i^2 n))}$. Let $\mathcal{C} = \{g \in \mathbb{R}^N \mid |g_i| \leq c_i \; \forall i \in [N]\}$. Following discussion in Section 2.6, for an adaptive regret bound of $B(i) + K$ to be achievable by a randomized algorithm such as Algorithm 6 we need

$$\mathcal{V}_n^{\text{oco}}([N], \mathcal{B}) := \left\langle\!\!\left\langle \inf_{P_t \in \Delta(\Delta_N)} \sup_{g_t \in \mathcal{C}} \mathbb{E}_{p_t \sim P_t} \mathbb{E}_{i_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^n \sup_{i \in [N]} \left[\sum_{t=1}^n \langle e_{i_t}, g_t \rangle - \sum_{t=1}^n \langle e_i, g_t \rangle - \mathcal{B}(i) \right] \leq K.$$

In the context of Algorithm 6, the distributions p_t above refer to the strategy $p_t(\sigma_{t+1:n})$ selected by the algorithm and P_t refers to the distribution over this strategy induced by sampling the random variables $\sigma_{t+1:n}$.

We will develop an algorithm to certify this bound for $K = 1$ using the framework of adaptive relaxations proposed by Foster et al. (2015). Define a relaxation $\mathbf{Rel} : \bigcup_{t=0}^n \mathcal{C}^t \to \mathbb{R}$ via

$$\mathbf{Rel}(g_{1:t}) := \mathbb{E}_{\sigma_{t+1:n} \in \{\pm 1\}^N} \sup_{i \in [N]} \left[-\sum_{s=1}^t \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^n \sigma_s[i] c_i - \mathcal{B}(i) \right].$$

The proof structure is as follows: We show that playing p_t as suggested by Algorithm 6 with $\mathbf{Rel}$ satisfies the initial condition and admissibility condition for adaptive relaxations from Foster et al. (2015), which implies that if we play p_t we will have $\mathrm{Reg}_n(i) \leq \mathcal{B}(i) + \mathbf{Rel}(\cdot)$. Then as a final step we bound $\mathbf{Rel}(\cdot)$ using a probabilistic maximal inequality, Lemma 16.

Initial condition This condition asks that the initial value of the relaxation $\mathbf{Rel}$ upper bound the worst-case value of the negative benchmark minus the bound $\mathcal{B}(i)$ (in other words, the inner part of $\mathcal{V}_n$ with the learner's loss removed). This is holds by definition and is trivial to verify:

$$\mathbf{Rel}(g_{1:n}) = \sup_{i \in [N]} \left[-\sum_{t=1}^{n} \langle e_i, g_t \rangle - \mathcal{B}(i) \right].$$

Admissibility For this step we must show that the inequality

$$\inf_{P_t \in \Delta(\Delta_N)} \sup_{g_t \in \mathcal{C}} \mathop{\mathbb{E}}_{p_t \sim P_t} \mathop{\mathbb{E}}_{i_t \sim p_t} [\langle e_{i_t}, g_t \rangle + \mathbf{Rel}(g_{1:t})] \leq \mathbf{Rel}(g_{1:t-1})$$

holds for each timestep t, and further that the inequality is certified by the strategy of Algorithm 6. We begin by expanding the definition of $\mathbf{Rel}$:

$$\inf_{P_t \in \Delta(\Delta_N)} \sup_{g_t \in \mathcal{C}} \mathop{\mathbb{E}}_{p_t \sim P_t} \mathop{\mathbb{E}}_{i_t \sim p_t} [\langle e_{i_t}, g_t \rangle + \mathbf{Rel}(g_{1:t})]$$

$$= \inf_{P_t \in \Delta(\Delta_N)} \sup_{g_t \in \mathcal{C}} \mathop{\mathbb{E}}_{p_t \sim P_t} \mathop{\mathbb{E}}_{i_t \sim p_t} \left[\langle e_{i_t}, g_t \rangle + \mathop{\mathbb{E}}_{\sigma_{t+1:n} \in \{\pm 1\}^N} \sup_{i \in [N]} \left[-\sum_{s=1}^{t} \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^{n} \sigma_s[i] c_i - \mathcal{B}(i) \right] \right].$$

Now plug in the randomized strategy given by Algorithm 6, with $\mathbb{E}_{\sigma_{t+1:n} \in \{\pm 1\}^N}$ taking the place of $\mathbb{E}_{p_t \sim P_t}$. This leads to an upper bound of

$$\sup_{g_t \in \mathcal{C}} \left[\mathop{\mathbb{E}}_{\sigma_{t+1:n} \in \{\pm 1\}^N} \left[\mathop{\mathbb{E}}_{i_t \sim p_t(\sigma_{t+1:n})} \langle e_{i_t}, g_t \rangle \right] + \mathop{\mathbb{E}}_{\sigma_{t+1:n}} \sup_{i \in [N]} \left[-\sum_{s=1}^{t} \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^{n} \sigma_s[i] c_i - \mathcal{B}(i) \right] \right].$$

Grouping expectations and applying Jensen's inequality:

$$\leq \mathop{\mathbb{E}}_{\sigma_{t+1:n} \in \{\pm 1\}^N} \sup_{g_t \in \mathcal{C}} \left[\mathop{\mathbb{E}}_{i_t \sim p_t(\sigma_{t+1:n})} \langle e_{i_t}, g_t \rangle + \sup_{i \in [N]} \left[-\sum_{s=1}^{t} \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^{n} \sigma_s[i] c_i - \mathcal{B}(i) \right] \right].$$

Expanding the definition of p_t (using its optimality in particular):

$$= \mathop{\mathbb{E}}_{\sigma_{t+1:n} \in \{\pm 1\}^N} \inf_{p_t \in \Delta_N} \sup_{g_t \in \mathcal{C}} \left[\langle p_t, g_t \rangle + \sup_{i \in [N]} \left[-\sum_{s=1}^{t} \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^{n} \sigma_s[i] c_i - \mathcal{B}(i) \right] \right].$$

We now apply a somewhat standard sequential symmetrization procedure. We begin by using the minimax theorem (Section 2.6) to swap the order of $\inf_{p_t}$ and $\sup_{g_t}$. To do so, we allow the g_t player to randomize, and denote their distribution by $Q_t \in \Delta(\mathcal{C})$.

$$= \mathop{\mathbb{E}}_{\sigma_{t+1:n} \in \{\pm 1\}^N} \sup_{Q_t \in \Delta(\mathcal{C})} \inf_{p_t \in \Delta_N} \mathop{\mathbb{E}}_{g_t \sim Q_t} \left[\langle p_t, g_t \rangle + \sup_{i \in [N]} \left[-\sum_{s=1}^{t} \langle e_i, g_s \rangle + 4 \sum_{s=t+1}^{n} \sigma_s[i] c_i - \mathcal{B}(i) \right] \right].$$

Since the supremum over i does not directly depend on p_t, we can rewrite this expression by introducing a (conditionally) i.i.d. copy of g_t which we will denote as g_t':

$$= \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{Q_t\in\Delta(\mathcal{C})} \mathop{\mathbb{E}}_{g_t\sim Q_t} \left[\sup_{i\in[N]} \left[\inf_{p_t\in\Delta_N} \mathop{\mathbb{E}}_{g_t'\sim Q_t} [\langle p_t, g_t'\rangle] - \sum_{s=1}^{t} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right]\right].$$

Choosing p_t to match e_i:

$$\leq \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{Q_t\in\Delta(\mathcal{C})} \mathop{\mathbb{E}}_{g_t\sim Q_t} \sup_{i\in[N]} \left[\mathop{\mathbb{E}}_{g_t'\sim Q_t} [\langle e_i, g_t'\rangle] - \langle e_i, g_t\rangle - \sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right].$$

Applying Jensen's inequality:

$$\leq \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{Q_t\in\Delta(\mathcal{C})} \mathop{\mathbb{E}}_{g_t,g_t'\sim Q_t} \sup_{i\in[N]} \left[\langle e_i, g_t'\rangle - \langle e_i, g_t\rangle - \sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right].$$

At this point we can introduce a new Rademacher random variable ϵ_t without changing the distribution of $g_t' - g_t$, thereby not changing the value of the game, then split the supremum:

$$= \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{Q_t\in\Delta(\mathcal{C})} \mathop{\mathbb{E}}_{\epsilon_t\in\{\pm1\}} \mathop{\mathbb{E}}_{g_t,g_t'\sim Q_t} \sup_{i\in[N]} \left[\epsilon_t\langle e_i, g_t' - g_t\rangle - \sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right]$$

$$\leq \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{Q_t\in\Delta(\mathcal{C})} \mathop{\mathbb{E}}_{\epsilon_t\in\{\pm1\}} \mathop{\mathbb{E}}_{g_t\sim Q_t} \sup_{i\in[N]} \left[2\epsilon_t\langle e_i, g_t\rangle - \sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right]$$

The above expression is now linear in Q_t, so it may be replaced with a pure strategy:

$$= \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{g_t\in\mathcal{C}} \mathop{\mathbb{E}}_{\epsilon_t\in\{\pm1\}} \sup_{i\in[N]} \left[2\epsilon_t\langle e_i, g_t\rangle - \sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right]$$

This expression is also convex in g_t, which means that the supremum will be obtained at a vertex of $\mathcal{C}$:

$$= \mathop{\mathbb{E}}_{\sigma_{t+1:n}\in\{\pm1\}^N} \sup_{\sigma_t\in\{\pm1\}^N} \mathop{\mathbb{E}}_{\epsilon_t\in\{\pm1\}} \sup_{i\in[N]} \left[2\epsilon_t\sigma_t[i]c_i - \sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right]$$

Now apply Theorem 31 conditioned on $\sigma_{t+1:n}$, with $w_i = -\sum_{s=1}^{t-1}\langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n}\sigma_s[i]c_i - \mathcal{B}(i)$.

$$\leq \mathop{\mathbb{E}}_{\sigma_{t:n}\in\{\pm1\}^N} \sup_{i\in[N]} \left[-\sum_{s=1}^{t-1} \langle e_i, g_s\rangle + 4\sum_{s=t}^{n} \sigma_s[i]c_i - \mathcal{B}(i) \right]$$

$$= \mathbf{Rel}(g_{1:t-1}).$$

Final value The final value of the relaxation is

$$\mathbf{Rel}(\cdot) = 2 \underset{\sigma_{1:n}\in\{\pm 1\}^N}{\mathbb{E}} \sup_{i\in[N]} \left[2\sum_{t=1}^{n} \sigma_t[i]c_i - 5c_i\sqrt{n(\log(1/\pi_i) + \log(4c_i^2 n))} \right] \leq 2\sum_{i\in[N]} \frac{\pi_i}{4c_i^2 n} \leq 1.$$

To show the first inequality we have applied a maximal inequality, Lemma 16, by recognizing that $\mathbf{Rel}(\cdot)$ is a supremum of a random process. Namely, we can write $\mathbf{Rel}(\cdot)$ in the form $\mathbb{E}\sup_{i\in[N]}\{X_i - B(i)\}$ with $X_i = 2\sum_{t=1}^{n}\sigma_t[i]c_i$. The standard mgf bound of $\mathbb{E}\, e^{\lambda X} \leq e^{\lambda^2(b-a)^2/8}$ for mean-zero random variables X with $a \leq X \leq b$ (Boucheron et al., 2013), along with independence of the Rademacher random variables in X_i, implies that X_i enjoys an mgf bound of

$$\mathbb{E}\, e^{\lambda X_i} \leq e^{2c_i^2\lambda^2 n}.$$

So to prove the result it suffices to take $h_i = 4c_i^2 n$ and $p = 2$ in the statement of Lemma 16 and note that $B(i) \geq (2 + 1/p)h_i^{1/p}(\log(h_i) + \log(1/\pi_i))^{1-1/p}$ in the notation of the lemma. The only additional detail to verify is that, since it was assumed that $c_i \geq 1$ for all i and since $n \geq 1$ by definition, the condition $h_i/\pi_i \geq e$ required by Lemma 16 is satisfied.

Computational efficiency We briefly sketch how the min-max optimization problem in the learner's strategy can be computed efficiently. Recall that the optimization problem is

$$\begin{aligned} &\min_{p\in\Delta_N} \sup_{g_t:|g_t[i]|\leq c_i} \left[\langle p, g_t\rangle + \sup_{i\in[N]} \left[-\sum_{s=1}^{t}\langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n}\sigma_s[i]c_i - B(i) \right]\right] \\ &= \min_{p\in\Delta_N} \sup_{i\in[N]} \sup_{g_t:|g_t[i]|\leq c_i} \left[\langle p, g_t\rangle - \sum_{s=1}^{t}\langle e_i, g_s\rangle + 4\sum_{s=t+1}^{n}\sigma_s[i]c_i - B(i) \right] \end{aligned}$$

Let $G_{t-1}(i) = \sum_{s=1}^{t-1} g_s[i]$. Since the quantity in the brackets above is linear in g_t and there are no interactions between coordinates, we can verify that conditioned on i the max over g_t is obtained via

$$\begin{aligned} &= \min_{p\in\Delta_N} \sup_{i\in[N]} \left[\langle p, c\rangle + (1 - 2p[i])c_i - G_{t-1}(i) + 4\sum_{s=t+1}^{n}\sigma_s[i]c_i - B(i) \right] \\ &= \min_{p\in\Delta_N} \sup_{i\in[N]} [\langle p, c\rangle + \langle a, e_i\rangle - 2\langle p, \mathrm{diag}(c)e_i\rangle], \end{aligned}$$

where $a[i] = c_i - G_{t-1}(i) + 4\sum_{s=t+1}^{n}\sigma_s[i]c_i - B(i)$. We can now employ a standard reduction from saddle point optimization to linear programming, i.e.

$$\begin{aligned} \text{minimize} \quad & \langle p, c\rangle + s \\ \text{subject to} \quad & s \geq \langle a, e_i\rangle - 2\langle p, \mathrm{diag}(c)e_i\rangle \quad \forall i. \\ & p \in \Delta_N. \end{aligned}$$

Assuming that $\min_i c_i \geq 1$, this linear program can be solved to accuracy ϵ by interior point methods (e.g. Renegar (1988)) in time $O(N^{3.5}\log(\epsilon^{-1}\max_i c_i))$ or by Mirror-Prox (Nemirovski, 2004) in time $O(N\epsilon^{-1}\max_i c_i)$. Since our rates scale as $\sqrt{n}$ we can set $\epsilon = 1/(\sqrt{n}\max_i c_i)$ to conclude the result.

As a final implementation detail, we remark that similar to the FTPL algorithm in Rakhlin et al. (2012) one can draw each perturbation $\sigma_t[i]$, from the distribution $\mathcal{N}(0,1)$ instead of using Rademacher random variables. This allows one to replace each sum $\sum_{s=t}^{n}\sigma_s[i]$ with a draw from $\mathcal{N}(0, n-t)$ and therefore avoid spending $O(n)$ time per step sampling perturbations. We have omitted the details because — for most values of c and N used in our applications, at least — the time required to solve the saddle point optimization problem dominates the runtime, not the time to sample perturbations.

$\square$

Theorem 31. *For any $w \in \mathbb{R}^N$, any $c \in \mathbb{R}_+^N$,*

$$\sup_{\sigma\in\{\pm1\}^N}\mathop{\mathbb{E}}_{\epsilon\in\{\pm1\}}\max_{i\in[N]}\{w_i + 2\epsilon\sigma_i c_i\} \leq \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N}\max_{i\in[N]}\{w_i + 4\sigma_i c_i\}. \tag{9.12}$$

Proof of Theorem 31. Fix any $\sigma \in \{\pm1\}^N$. Let $i_1 = \arg\max_{i\in[N]}\{w_i + 2\sigma_i c_i\}$ and $i_{-1} =$

$\arg\max_{i\in[N]}\{w_i - 2\sigma_i c_i\}$. Then it is easy to see that

$$\begin{aligned}\mathop{\mathbb{E}}_{\epsilon}\max_{i\in[N]}\{w_i + 2\epsilon\sigma_i c_i\} = \mathop{\mathbb{E}}_{\epsilon}\max_{i\in\{i_1,i_{-1}\}}\{w_i + 2\epsilon\sigma_i c_i\} &\le \mathop{\mathbb{E}}_{\sigma'\in\{\pm1\}^N}\max_{i\in\{i_1,i_{-1}\}}\{w_i + 4\sigma_i' c_i\} \\ &\le \mathop{\mathbb{E}}_{\sigma'\in\{\pm1\}^N}\max_{i\in[N]}\{w_i + 4\sigma_i' c_i\}.\end{aligned}$$

The central inequality above follows by Lemma 17 with the pair $(w, 2c)$. Since the above bound holds for any σ, we conclude that (9.12) holds.

□

Lemma 17. For any pair (w, c) where $w \in \mathbb{R}^N$ any $c \in \mathbb{R}^N_+$, the inequality

$$\sup_{\sigma\in\{\pm1\}^N}\mathop{\mathbb{E}}_{\epsilon\in\{\pm1\}}\max_{i\in[N]}\{w_i + \epsilon\sigma_i c_i\} \le \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N}\max_{i\in[N]}\{w_i + 2\sigma_i c_i\}. \tag{9.13}$$

holds when $N = 2$.

Proof of Lemma 17. In this proof we adopt the notation that for any element $j \in [2]$, $-j$ denote the other element. Say the pair (w, c) is *dominated* if there exists j for which $w_j - c_j \ge w_{-j} + c_{-j}$. Note that this of course implies $w_j + c_j \ge w_{-j} + c_{-j}$ as well, since c is non-negative.

Dominated case Suppose (w, c) is dominated by index j. Then (9.13) holds trivially for any $K \in \mathbb{R}$ by

$$\sup_{\sigma\in\{\pm1\}^N}\mathop{\mathbb{E}}_{\epsilon}\max_{i\in[N]}\{w_i + \epsilon\sigma_i c_i\} = w_j = \max_{i\in[N]}\{w_i + K\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N}\sigma_i c_i\} \le \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N}\max_{i\in[N]}\{w_i + K\sigma_i c_i\}.$$

We now focus on the trickier "not dominated" case.

Rescaling doesn't induce domination We first observe that if (w, c) does is not dominated, (w, Bc) is not dominated either for any $B \ge 1$. Let j be the index for which

$w_j + c_j \geq w_{-j} + c_{-j}$ which implies $w_j - c_j \leq w_{-j} + c_{-j}$ because (w, c) is not dominated. Observe that if (w, Bc) is dominated we either have $w_j - Bc_j \geq w_{-j} + Bc_{-j}$ or $w_{-j} - Bc_{-j} \geq w_j + Bc_j$. The first case cannot hold because $B \geq 1$ and we already know that (w, c) is not dominated. The second case in particular implies $w_{-j} \geq w_j$, so we must have had $c_j \geq c_{-j}$ to begin with. But in that case we will still have $w_j + Bc_j \geq w_{-j} + Bc_{-j}$ which contradicts the domination.

Note: It is good to keep in mind that while rescaling does not induce domination, it may not be the case in general that $w_j + Bc_j \geq w_{-j} + Bc_{-j}$ even though $w_j + c_j \geq w_{-j} + c_{-j}$. That is, the "leader" may change after rescaling.

LHS of (9.13) for (w, c) not dominated When (w, c) is not dominated we have

$$\sup_{\sigma \in \{\pm 1\}^N} \mathbb{E}_{\epsilon} \max_{i \in [N]} \{w_i + \epsilon \sigma_i c_i\} = \frac{1}{2}(w_1 + c_1) + \frac{1}{2}(w_2 + c_2).$$

RHS of (9.13) for (w, c) not dominated We will consider the RHS of (9.13) for $(w, c') := (w, Bc)$ for some $B \geq 1$ to be decided. By the argument above, the pair (w, c') is also not dominated. For the remainder of the proof, 1 will denote the index for which $w_1 + c'_1 \geq w_2 + c'_2$. Because the pair is not dominated, the value the RHS takes can be classified into two cases based on the relationship between c' and w.

- Case 1: $w_1 - c'_1 \leq w_2 - c'_2$:

  In this case there is equal probability that the process takes on value $w_2 - c'_2$ or $w_2 + c'_2$ conditioned on the event that $\sigma_1 = -1$, so we have the equality:

$$\mathbb{E}_{\sigma \in \{\pm 1\}^N} \max_{i \in [N]} \{w_i + \sigma_i c'_i\} = \frac{1}{2}(w_1 + w_2) + \frac{1}{2} c'_1$$

Furthermore, Case 1 implies $c'_1 \geq c'_2$, which leads to an inequality:

$$\geq \frac{1}{2}(w_1 + w_2) + \frac{1}{4}(c'_1 + c'_2).$$

- Case 2: $w_1 - c'_1 \geq w_2 - c'_2$:

  In this case, conditioned on the event that $\sigma_1 = -1$, there is equal probability that the process takes on value $w_2 + c'_2$ or $w_1 - c'_1$, so the equality becomes:

  $$\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N} \max_{i\in[N]}\{w_i + \sigma_i c'_i\} = \frac{1}{2}(w_1 + c'_1) + \frac{1}{4}(w_2 + c'_2) + \frac{1}{4}(w_1 - c'_1)$$

  Case 2 implies that $w_1 \geq w_2$, because we may add the inequalities $w_1 + c'_1 \geq w_2 + c'_2$ and $w_1 - c'_1 \geq w_2 - c'_2$. This gives an inequality:

  $$\geq \frac{1}{2}(w_1 + w_2) + \frac{1}{4}(c'_1 + c'_2).$$

Combining our results for the two cases, we have that for any vector c', so long as (w, c') is not dominated,

$$\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N} \max_{i\in[N]}\{w_i + \sigma_i c'_i\} \geq \frac{1}{2}(w_1 + w_2) + \frac{1}{4}(c'_1 + c'_2).$$

In particular, choosing $B = 2$ implies (9.13) in the non-dominated case:

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N} \max_{i\in[N]}\{w_i + 2\sigma_i c_i\} &\geq \frac{1}{2}(w_1 + w_2) + \frac{1}{2}(c_1 + c_2)\\ &= \sup_{\sigma\in\{\pm1\}^N} \mathop{\mathbb{E}}_{\epsilon} \max_{i\in[N]}\{w_i + \epsilon\sigma_i c_i\}.\end{aligned}$$

Final result Combining the dominated and non-dominated results we have that for any (w, c).

$$\sup_{\sigma\in\{\pm1\}^N} \mathop{\mathbb{E}}_{\epsilon} \max_{i\in[N]}\{w_i + \epsilon\sigma_i c_i\} \leq \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^N} \max_{i\in[N]}\{w_i + 2\sigma_i c_i\}.$$

$\square$

Proof of Lemma 16. Let $B(i) = Ch_i^{1/p}(\log(h_i) + \log(1/\pi_i))^{1-1/p}$ for some constant C to be decided later. One should verify that $\log(h_i) + \log(1/\pi_i)$ is always non-negative by the assumption that $h_i/\pi_i \geq e$, which will be used repeatedly. To begin, observe that

$$\mathbb{E} \sup_{i \in [N]} \{X_i - B(i)\} \leq \mathbb{E} \sup_{i \in [N]} [X_i - B(i)]_+,$$

where $[x]_+ = \max\{x, 0\}$. By non-negativity of $[x]_+$ it further holds that

$$\leq \mathbb{E} \sum_{i \in [N]} [X_i - B(i)]_+.$$

Fixing an arbitrary sequence $(\lambda_i)_{i \in [N]}$ with $\lambda_i > 0$, the basic inequality $\max\{a, b\} \leq \frac{1}{\lambda} \log(e^{\lambda a} + e^{\lambda b})$ implies the following upper bound:

$$\leq \mathbb{E} \sum_{i \in [N]} \frac{1}{\lambda_i} \log\bigl(1 + e^{\lambda_i(X_i - B(i))}\bigr).$$

Apply Jensen's inequality:

$$\leq \sum_{i \in [N]} \frac{1}{\lambda_i} \log\bigl(1 + \mathbb{E}\, e^{\lambda_i(X_i - B(i))}\bigr).$$

Now use the moment bound assumed in the lemma statement:

$$\leq \sum_{i \in [N]} \frac{1}{\lambda_i} \log\Bigl(1 + e^{\left(\lambda_i^p h_i - \lambda_i B(i)\right)}\Bigr).$$

Lastly, apply the inequality $\log(1 + x) \leq x$ for $x \geq 0$:

$$\leq \sum_{i \in [N]} \exp(\lambda_i^p h_i - \lambda_i B(i) + \log(1/\lambda_i)).$$

We now take $\lambda_i = \left(\frac{\log(h_i) + \log(1/\pi_i)}{h_i}\right)^{1/p}$ and bound each exponent in the sum above. Using the definition of $B(i)$:

$$\lambda_i^p h_i - \lambda_i B(i) + \log(1/\lambda_i) = \log(1/\lambda_i) - (C - 1)(\log(1/\pi_i) + \log(h_i)).$$

Next observe that

$$\log(1/\lambda_i) = \frac{1}{p} \log\left(\frac{h_i}{\log(h_i/\pi_i)}\right) \leq \frac{1}{p} \log(h_i),$$

where we have used that $h_i/\pi_i \geq e$. With this, and using that $\log(1/\pi_i) \geq 0$, we have

$$\lambda_i^p h_i - \lambda_i B(i) + \log(1/\lambda_i) \leq \; -(C - 1 - 1/p)(\log(1/\pi_i) + \log(h_i)).$$

Taking $C \geq 2 + 1/p$ and using this bound in the summation over i yields the result:

$$\mathbb{E} \sup_{i \in [N]} \{X_i - B(i)\} \leq \sum_{i \in [N]} \frac{\pi_i}{h_i}.$$

□

9.3.2 Proofs for Section 9.2.2

Proof of Theorem 23. First, we verify that the loss sequence $(g_t)_{t \leq n}$ is such that the regret bound derived for MULTISCALEFTPL applies. In particular, we need to verify that $|g_t[i]| \leq c_i$ for each i. To this end, fix an index $i \in [N]$, and note that since f_t is L_i-Lipschitz on $\mathcal{W}_i$ with respect to the norm $\|\cdot\|_{(i)}$ we have

$$|g_t[i]| = |f_t(w_t^i) - f_t(0)| \leq L_i \left\| w_t^i - 0 \right\|_{(i)} \leq L_i R_i \leq L_i R_i = c_i,$$

as required. Also, it was assumed that $c_i = L_i R_i \geq 1$, as required for Theorem 22.

Now, recall that (p_t) is the sequence of distributions produced by the meta-algorithm. The algorithm's total loss with respect to the centered iterates $(\widetilde{f}_t)$ is given by

$$\sum_{t=1}^{n} \widetilde{f}_t(w_t^{i_t}) = \sum_{t=1}^{n} \langle e_{i_t}, g_t \rangle,$$

where this equality is due to the construction of the losses $(g_t)_{t \leq n}$ given to MULTI-SCALEFTPL. The regret bound for MULTISCALEFTPL now implies that

$$\mathbb{E}\left[\sum_{t=1}^{n} \langle e_{i_t}, g_t \rangle - \min_{i \in [N]} \left\{ \sum_{t=1}^{n} g_t[i] + O\Big(R_i L_i \sqrt{n \log(R_i L_i n/\pi_i)}\Big) \right\}\right] \leq 0,$$

where we have obtained this inequality by substituting the value of the vector c constructed by MULTISCALEOCO into the regret bound (9.3) for MULTISCALEFTPL. Now, observe that for each i we have

$$\sum_{t=1}^{n} g_t[i] = \sum_{t=1}^{n} \widetilde{f}_t(w_t^i) \leq \inf_{w\in\mathcal{W}_i} \sum_{t=1}^{n} \widetilde{f}_t(w) + \mathrm{Reg}_n(i),$$

where we have used the definition of g_t and the regret bound assumed on the sub-algorithm. Combining these inequalities, we have

$$\mathbb{E}\left[\sum_{t=1}^{n} \widetilde{f}_t(w_t^{i_t}) - \min_{i\in[N]}\left\{\inf_{w\in\mathcal{W}_i} \sum_{t=1}^{n} \widetilde{f}_t(w) + \mathrm{Reg}_n(i) + O\Big(R_iL_i\sqrt{n\log(R_iL_in/\pi_i)}\Big)\right\}\right] \leq 0.$$

Finally, observe that since $\widetilde{f}_t(w) = f_t(w) - f_t(0)$, the above is equivalent to

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{i_t}) - \min_{i\in[N]}\left\{\inf_{w\in\mathcal{W}_i} \sum_{t=1}^{n} f_t(w) + \mathrm{Reg}_n(i) + O\Big(R_iL_i\sqrt{n\log(R_iL_in/\pi_i)}\Big)\right\}\right] \leq 0.$$

□

Mirror Descent Online Mirror Descent is the standard algorithm for online linear optimization over convex sets. It is parameterized by a convex set $\mathcal{W}$, learning rate η, and strongly convex regularizer $\mathcal{R} : \mathcal{W} \to \mathbb{R}$. We define the update MIRRORDESCENT$(\eta, \mathcal{W}, \mathcal{R})$ as follows.

First, set $w_1 = \arg\min_{w\in\mathcal{W}} \mathcal{R}(w)$. Then, for each time $t \in [n]$:

- Receive gradient g_t and let $\widetilde{w}_{t+1}$ satisfy $\nabla\mathcal{R}(\widetilde{w}_{t+1}) = \nabla\mathcal{R}(w_t) - \eta g_t$.
- Set $w_{t+1} = \arg\min_{w\in\mathcal{W}} \mathcal{D}_{\mathcal{R}}(w \mid \widetilde{w}_{t+1})$.

Fact 1 (Mirror Descent (e.g. Hazan (2016))). Let (w_t) be the iterates produced by MIRRORDESCENT$(\eta, \mathcal{W}, \mathcal{R})$ on a sequence of vectors $(g_t)_{t\leq n}$. If $\mathcal{R}$ is λ-strongly convex with respect to a norm $\|\cdot\|_{\mathcal{R}}$, the iterates satisfy

$$\sum_{t=1}^{n}\langle w_t - w, g_t\rangle \leq \frac{\eta}{2\lambda}\sum_{t=1}^{n}\|g_t\|_{\mathcal{R},\star}^2 + \frac{1}{\eta}\mathcal{R}(w) \quad \forall w \in \mathcal{W}. \tag{9.14}$$

Proof of Theorem 24. Recall that each sub-algorithm $\textsc{Alg}_i$ runs Mirror Descent over a ball in $(\mathfrak{B}, \|\cdot\|)$ of radius R_i using the regularizer $\mathcal{R}(w) = \frac{1}{2}\|w\|^2$. From the regret bound for Mirror Descent (1), the meta-algorithm's choice of Mirror Descent parameters for $\textsc{Alg}_i$ (in particular, the choice $\eta_i = \frac{R_i}{L}\sqrt{\frac{\lambda}{n}}$) guarantees that

$$\sum_{t=1}^{n} f_t(w_t^i) - \inf_{w \in \mathcal{W}_i} \sum_{t=1}^{n} f_t(w) \le O(R_i L \sqrt{n/\lambda}).$$

Combined with the regret bound for $\textsc{MultiScaleOCO}$ (Theorem 23, noting that $R_i L_i = R_i L \ge 1$), this implies that the meta-algorithm's regret satisfies

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{i_t}) - \min_{i \in [N]}\left\{\inf_{w \in \mathcal{W}_i} \sum_{t=1}^{n} f_t(w) + O(R_i L\sqrt{n/\lambda}) + O\Big(R_i L\sqrt{n\log(R_i L n/\pi_i)}\Big)\right\}\right] \le 0.$$

Which, using that $\pi_i = 1/(n+1)$ and combining terms, further implies

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{i_t}) - \min_{i \in [N]}\left\{\inf_{w \in \mathcal{W}_i} \sum_{t=1}^{n} f_t(w) + O\Big(R_i L\sqrt{n\log(R_i L n)/\lambda}\Big)\right\}\right] \le 0.$$

Now, recall that $i \in [n+1]$, and that $R_i = e^{i-1}$. Consider the algorithm's regret against a comparator w. For now, assume that w satisfies $1 \le \|w\| \le e^n$ — we will see shortly that this is without loss of generality. Let $i^\star(w) = \min\{i \mid w \in \mathcal{W}_i\}$. Then the regret bound above implies

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{i_t}) - \left\{\sum_{t=1}^{n} f_t(w) + O\left(R_{i^\star(w)} L\sqrt{n\log\big(R_{i^\star(w)} L n\big)/\lambda}\right)\right\}\right] \le 0.$$

Furthermore, since $R_i = e^{i-1}$, we have that $R_{i^\star(w)} \le e\|w\|$, and so

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{i_t}) - \left\{\sum_{t=1}^{n} f_t(w) + O\Big(\|w\| L\sqrt{n\log(\|w\| L n/)\lambda}\Big)\right\}\right] \le 0.$$

This is exactly the regret bound we wanted. Now, the case where $\|w\| \le 1$ is handled by simply noting $i^\star(w) = 1$ and writing $R_1 = 1 \le 1 + \|w\|$, which gives the $\|w\| + 1$ factor as follows:

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{i_t}) - \left\{\sum_{t=1}^{n} f_t(w) + O\Big((\|w\|+1) L\sqrt{n\log((\|w\|+1) L n/)\lambda}\Big)\right\}\right] \le 0.$$

To handle the case where $\|w\| \geq e^n$ we appeal to Corollary 12 with $c = L\sqrt{n}$ and $\gamma = 1/2$, which shows that it suffices to consider only $\|w\| \leq \exp\Big(\big(\frac{Ln}{c}\big)^{1/\gamma}\Big) = e^n$. Note that the constants appearing in the regret bound above, both inside the $O(\cdot)$ and inside the $\sqrt{\log(\cdot)}$ are worse than those with which we instantiate Corollary 12. This is not an issue because worse constants only reduce the radius that must be considered in the corollary. $\square$

Lemma 18. Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be given. Suppose the loss sequence $(f_t)_{t\leq n}$ is L-Lipschitz with respect to $\|\cdot\|_\star$. Then a regret bound of the form

$$\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w) \leq F(\|w\|) \quad \forall w \in \mathfrak{B} \tag{9.15}$$

holds if the restricted regret bound

$$\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w) \leq F(\|w\|) \quad \forall f : \|f\| \leq \alpha^\star, \tag{9.16}$$

holds, where $\alpha^\star$ is the greatest non-negative number for which $F(\alpha^\star) - \alpha^\star Ln \geq F(0)$.

Proof of Lemma 18. Assume wlog that $f_t(0) = 0$ for each t. This is possible because

$$\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w) = \sum_{t=1}^{n} (f_t(w_t) - f_t(0)) - \sum_{t=1}^{n} (f_t(w) - f_t(0)).$$

To begin, observe that (9.15) is equivalent to

$$\sum_{t=1}^{n} f_t(w_t) \leq \inf_{w\in\mathfrak{B}} \left\{ \sum_{t=1}^{n} f_t(w) + F(\|w\|) \right\}.$$

By selecting $w = 0$, $f_t(0) = 0$ implies that the infimum on the right is always upper bounded in value by $F(0)$. In the other direction, Lipschitzness of the losses along with $f_t(0) = 0$ implies that the infimum is lower bounded as

$$\inf_{w\in\mathfrak{B}} \left\{ \sum_{t=1}^{n} f_t(w) + F(\|w\|) \right\} \geq \inf_{w\in\mathfrak{B}} \{-L\|w\|n + F(\|w\|)\} = \inf_{\alpha\geq 0} \{-\alpha Ln + F(\alpha)\}.$$

Therefore if $\alpha \geq \alpha^\star$, the lower bound $-\alpha Ln + F(\alpha)$ will be sub-optimal compared to the upper bound of $F(0)$ obtained by choosing $\alpha = 0$. $\square$

Corollary 12. When $F(r) = c \cdot (r+1)\log(r+1)^{\gamma}$ for $\gamma > 0$, it is sufficient to consider

$$\sum_{t=1}^{n} f_t(w_t) - \sum_{t=1}^{n} f_t(w) \leq F(\|w\|) \quad \forall w : \|w\| \leq \exp\left(\left(\frac{Ln}{c}\right)^{1/\gamma}\right). \tag{9.17}$$

Proof of Corollary 12. Note that $F(0) = 0$. Let r denote the minimizer of $F(\alpha) - \alpha \cdot a$ (where $a = Ln$). Differentiating this expression yields

$$a = c\big(\log(r+1)^{\gamma} + \gamma\log(r+1)^{\gamma-1}\big),$$

which further implies

$$\log(r+1)^{\gamma} = \frac{a}{c} \cdot \frac{1}{1+\gamma/\log(r+1)} \leq \frac{a}{c}.$$

Rearranging, we have $r \leq \exp((a/c)^{1/\gamma}) - 1$. Since $F(\alpha) - \alpha \cdot a$ is strictly convex, this function is increasing above r. To conclude, we guess an upper bound on the value of $\alpha^{\star}$: $\alpha := \exp((a/c)^{1/\gamma}) - 1$. Substituting this value in, we have

$$F(\alpha) - \alpha \cdot a \geq a\exp((a/c)^{1/\gamma}) - a \cdot \exp((a/c)^{1/\gamma}) = 0 = F(0),$$

which yields the result. $\square$

Proof of Theorem 25. We only sketch the details of this proof as it follows Theorem 24 very closely.

We first describe sub-algorithm configuration for MULTISCALEOCO that achieves the claimed regret bound. Our strategy will be to take a discretization the range of p values $[1+\delta, 2]$, and produce a set of sub-algorithms for each p in this discrete set. For a fixed p, the construction of the set of sub-algorithms will be exactly is in Theorem 24. The discrete set of ps will have the form $p_k = 1 + \delta + \min\{(k-1)\cdot\epsilon, (1-\delta)\}$, for $\epsilon = 1/\log(d)$ and $k \in [1, \ldots, K]$, where $K = \lceil (1-\delta)/\epsilon \rceil + 1$ (in particular $k \leq \log(d) + 1$).

For a fixed k, the norm $\|\cdot\|_{p_k}$ has that $\frac{1}{2}\|\cdot\|_{p_k}^2$ is $(p_k - 1)$-strongly convex with respect to itself (Kakade et al., 2009a). With this in mind, we create a set of $N := K(n+1)$ sub-algorithms, which we will index by pairs $(k, j) \in [K] \times [n+1]$ instead of $i \in [K(n+1)]$ for notational convenience.

- For each $k \in [K]$:
  - $L_k = L_{p_k}$.
  - For each $j \in \{1, \ldots, n+1\}$:
    * Set $R_j = e^{j-1}$.
    * Take $\mathcal{W}_{(k,j)} = \left\{w \in \mathfrak{B} \mid \|w\|_{p_k} \le R_j\right\}$, $\eta_{(k,j)} = \frac{R_j}{L_k}\sqrt{\frac{\lambda_{p_k}}{n}}$, where $\lambda_{p_k} = (p_k - 1)$.
    * Let $\textsc{Alg}_j = \textsc{MirrorDescent}(\eta_{(k,j)}, \mathcal{W}_{(k,j)}, \|\cdot\|_{p_k}^2)$.
- $\pi = \mathrm{Uniform}([K] \times [n+1])$.

Clearly the total number of sub-algorithms and hence the running time scales as $O(n \cdot \log(d))$.

Referring back to the proof of Theorem 24, and letting (k_t, j_t) denote the index pair chosen by $\textsc{MultiScaleOCO}$ in round t, it is clear that for a fixed k, the algorithm satisfies for all $w \in \mathbb{R}^d$

$$\mathbb{E}\left[\sum_{t=1}^n f_t(w_t^{(k_t,j_t)}) - \left\{\sum_{t=1}^n f_t(w) + O\left((\|w\|_{p_k}+1)L_{p_k}\sqrt{n\log((\|w\|_{p_k}+1)L_{p_k} n\log(d))/(p_k-1)}\right)\right\}\right] \le 0.$$

In fact, the regret guarantee for $\textsc{MultiScaleOCO}$ implies that

$$\mathbb{E}\left[\sum_{t=1}^n f_t(w_t^{(k_t,j_t)}) - \min_{k\in[N]}\left\{\sum_{t=1}^n f_t(w) + O\left((\|w\|_{p_k}+1)L_{p_k}\sqrt{n\log((\|w\|_{p_k}+1)L_{p_k} n\log(d))/(p_k-1)}\right)\right\}\right] \le 0.$$

We now appeal to the choice of discretization to deduce that

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{(k_t,j_t)}) - \min_{p\in[1+\delta,2]}\left\{\sum_{t=1}^{n} f_t(w) + O\left((\|w\|_p+1)L_p\sqrt{n\log((\|w\|_p+1)L_p\log(d)n)/(p-1)}\right)\right\}\right] \leq 0.$$

Suppose there is some $p \in [1+\delta, 2]$ of interest. Let k be the greatest integer for which $p_k \leq p$. We claim that the bound

$$\mathbb{E}\left[\sum_{t=1}^{n} f_t(w_t^{(k_t,j_t)}) - \left\{\sum_{t=1}^{n} f_t(w) + O\left((\|w\|_{p_k}+1)L_{p_k}\sqrt{n\log((\|w\|_{p_k}+1)L_{p_k}n\log(d))/(p_k-1)}\right)\right\}\right] \leq 0,$$

implies the desired result. By duality we have that $\|w\|_{p_k} \geq \|w\|_p$ and $L_{p_k} \leq L_p$. To conclude, observe that $\|w\|_{p_k}/\|w\|_p \leq \|w\|_{p_k}/\|w\|_{p_{k+1}} \leq d^{\epsilon} = d^{1/\log(d)} = O(1)$, so the norm terms in the bound above are within constant factors of the desired bound. □

Proof of Theorem 26. Recall that for fixed k, the learner predicts from a class

$$\mathcal{W}_k = \left\{W \in \mathbb{R}^{d\times d} \mid W \succeq 0, \|W\|_\sigma \leq 1, \langle W, I\rangle = k\right\},$$

and experiences affine losses $f_t(W_t) = \langle I - W_t, Y_t\rangle$, where $Y_t \in \mathcal{Y} := \left\{Y \in \mathbb{R}^{d\times d} \mid Y \succeq 0, \|Y\|_\sigma \leq 1\right\}$.

The regret for this game is given by

$$\sup_{W\in\mathcal{W}_k}\left[\sum_{t=1}^{n}\langle I - W_t, Y_t\rangle - \sum_{t=1}^{n}\langle I - W, Y_t\rangle\right]. \tag{9.18}$$

From Nie et al. (2013), we have that for fixed k the strategy Matrix Exponentiated Gradient has regret bounded by

$$O\left(\min\left\{\sqrt{nk^2\log(n/k)}, \sqrt{n(d-k)^2\log(n/(d-k))}\right\}\right) = \widetilde{O}\left(\sqrt{n\min\{k, d-k\}^2}\right).$$

Note: The variant of Matrix Exponentiated Gradient that obtains this strategy uses either losses or gains depending on the value of k. See Nie et al. (2013) for more details.

The configuration with which we invoke MULTISCALEOCO is:

- For each $i \in [\lceil \log(d/2) \rceil + 1]$:
  - Set $R_i = e^{i-1}$, $L_i = 1$.
  - $\mathcal{W}_i = \left\{ W \in \mathbb{R}^{d\times d} \mid W \succeq 0, \|W\|_\sigma \le 1, \langle W, I \rangle = R_i \right\}$
  - Take $\textsc{Alg}_i = \textsc{Matrix Exponentiated Gradient}(\mathcal{W}_i)$ as described in Nie et al. (2013).
- $\pi = \text{Uniform}([\lceil \log(d/2) \rceil + 1])$.

As in Theorem 24 and Theorem 25, choosing R_i to be spaced exponentially is sufficient to guarantee that there is a sub-algorithm whose regret is within a constant factor e of $\widetilde{O}(k\sqrt{n})$ for any choice of the rank k.

All that remains is that the losses of the sub-algorithms satisfy the claimed upper bound R_i. Observe that MULTISCALEOCO works with centered loss $\widetilde{f}_t(W) = -\langle W, Y_t \rangle$. For any $W \in \mathcal{W}_k$, we have

$$|\langle W, Y_t \rangle| \le \|Y_t\|_\sigma \|W\|_\Sigma \le 1 \cdot R_k,$$

so the condition is satisfied. □

Proof of Theorem 27. We will use a meta-algorithm strategy closely resembling that of the smooth Banach space setting. The only difference is that $\|\cdot\|_\Sigma$ is not smooth, so MATRIX MULTIPLICATIVE WEIGHTS, which uses the log-trace-exponential function as a surrogate for $\|\cdot\|_\Sigma$, is used as the sub-algorithm instead of working with $\|\cdot\|_\Sigma$ directly.

We use the version of MATRIX MULTIPLICATIVE WEIGHTS stated in Hazan et al. (2012) Theorem 13, which uses classes of the form $\mathcal{W}_r = \left\{ W \in \mathbb{R}^{d\times d} \mid W \succeq 0, \|W\|_\Sigma \le r \right\}$ and has regret against $\mathcal{W}_r$ bounded by $O(r\sqrt{n \log d})$ whenever each loss matrix Y_t has $\|Y_t\|_\sigma \leq 1$. Using this strategy for fixed r as a sub-algorithm for MULTISCALEOCO, we achieve the following oracle inequality efficiently:

For each $i \in [n+1]$:

- Set $R_i = 2^{i-1}$
- $L_i = 1$ (we are assuming $\|Y_t\|_\sigma \leq 1$).
- $\mathcal{W}_i = \left\{ W \in \mathbb{R}^{d \times d} \mid W \succeq 0, \|W\|_\Sigma \leq R_i \right\}$
- $\textsc{Alg}_i = \textsc{Matrix Multiplicative Weights}(\mathcal{W}_i)$

Finally, we set $\pi = \mathrm{Uniform}([n+1])$. That this configuration is sufficient follows from the doubling analysis given in the proof of Theorem 24. Losses are once again bounded via $|\langle W, Y_t \rangle| \leq \|W\|_\Sigma \|Y_t\|_\sigma \leq R_i$ for $W \in \mathcal{W}_i$. □

9.3.3 Proofs from Section 9.2.3

Proof of Theorem 28. This theorem is an immediate consequence of Theorem 23, using the absolute value $|\cdot|$ as the norm. The only significant detail one must check is that the proof of Theorem 23 uses the regret statement for each sub-algorithm as a black box, and so the nonlinearity of the comparator $\mathcal{F}$ does not change the analysis. □

Proof of Theorem 29. This is a corollary of Theorem 28. That theorem, configured with one sub-algorithm for each class $\mathcal{F}_k$ and with $L_k = L$, $R_k = R_k$, and $\pi_k = 1/k^2$, implies

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell(\widehat{y}_t^i, y_t) - \inf_{f \in \mathcal{F}_k} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \leq \mathbb{E}[\mathbf{Rad}_n(\mathcal{F}_k)] + O\Big(R_k L \sqrt{n \log(R_k L n k)}\Big) \quad \forall i \in [N]. \tag{9.19}$$

The final regret bounded stated follows from the assumed growth rate on $\mathbf{Rad}(\mathcal{F}_k)$. $\square$

Proof of Theorem 30. We briefly sketch the construction as follows:

1. For each $\mathcal{H}_k$, construct a sequence of nested subclasses (norm balls) as precisely as in the proof of Theorem 24. There will be $O(n)$ sub-algorithms for each such class.
2. For each sub-algorithm in class k, take the prior weight π proportional to $1/nk^2$.

Using the analysis from Theorem 24 — namely that for each norm $\|\cdot\|_{\mathcal{H}_k}$ it is sufficient to only consider predictors with norm bounded by e^n — , one can see that the result follows from Theorem 28. $\square$

9.4 Chapter Notes

This chapter is based on Foster et al. (2017a). For the special case of model selection in Banach spaces, faster algorithms were subsequently presented in Cutkosky and Orabona (2018) and Foster et al. (2018c).

Bubeck et al. (2017) simultaneously developed a multi-scale experts algorithm which could also be used in our framework. Their regret bound has sub-optimal dependence on the prior distribution over experts, but their algorithm is more efficient and is able to obtain multiplicative regret guarantees.

Detailed Discussion of Related Work There are two directions in parameter-free online learning that have been explored extensively. The first considers bounds of the form (9.2); namely, the Hilbert space version of the more general setting explored in Section 9.2.2.

Beginning with Mcmahan and Streeter (2012), which obtained a slightly looser rate than (9.2), research has focused on obtaining tighter dependence on $\|w\|_2$ and $\log(n)$ in this type of bound (McMahan and Abernethy, 2013; McMahan and Orabona, 2014; Orabona, 2014; Orabona and Pál, 2016); all of these algorithms run in linear time per update step. Cutkosky and Boahen (2016, 2017) extended these results to the case where the Lipschitz constant is not known in advance. These works give lower bounds for general norms, but only give efficient algorithms for Hilbert spaces.

The second direction concerns so-called "quantile bounds" (Chaudhuri et al., 2009; Koolen and van Erven, 2015; Luo and Schapire, 2015; Orabona and Pál, 2016) for experts setting, where the learner's decision set $\mathcal{W}$ is the simplex Δ_d and losses are bounded in ℓ_∞. The multi-scale machinery developed in the present work is not needed to obtain bounds for this setting because the losses are uniformly bounded across all model classes. Indeed, Foster et al. (2015) recovered a basic form of quantile bound using the vanilla multiplicative weights strategy as a meta-algorithm. It is not known whether the more sophisticated data-dependent quantile bounds given in Koolen and van Erven (2015); Luo and Schapire (2015) can be recovered in the same fashion.

CHAPTER 10

LOGISTIC REGRESSION, CLASSIFICATION, AND BOOSTING

This chapter addresses adaptivity to misspecification in statistical learning, and provides new guarantees for the fundamental statistical task of logistic regression. Logistic regression was originally introduced for binary classification in a *well-specified* statistical model where conditional class probabilities are realized by the logistic function (Cox, 1958). Agnostic learning guarantees for this setting imply generalization even when this assumption does not hold, adapting between the "nice" case where the model is well-specified and the purely noisy setting. Unfortunately, fast rates for learning linear predictors in this setting exhibit exponential dependence on the predictor norm, and Hazan et al. (2014) showed that this is unimprovable.

Starting with the simple observation that the logistic loss is 1-mixable, we design a new efficient *improper* learning algorithm for logistic regression that circumvents the aforementioned lower bound with a regret bound exhibiting a *doubly-exponential* improvement in dependence on the predictor norm. This provides a positive resolution to a variant of the COLT 2012 open problem of McMahan and Streeter (2012) when improper learning is allowed. This improvement is obtained both in the online setting and, with some extra work, in the batch statistical setting with high probability. We also show that the improved dependence on predictor norm is near-optimal, and use the equivalence of online prediction and martingale inequalities developed in Part II to give information-theoretic bounds on the optimal rates for improper logistic regression with general function classes. This characterizes the extent to which our improvement for linear classes extends to other parametric and even nonparametric settings.

Beyond the statistical learning setting, the improved logistic regression algorithm we develop yields the following applications: (a) we give algorithms for online bandit multiclass learning with the logistic loss with an $\tilde{O}(\sqrt{n})$ relative mistake bound across essentially all parameter ranges, thus providing a solution to the COLT 2009 open problem of Abernethy and Rakhlin (2009), and (b) we give an adaptive algorithm for online multiclass boosting with optimal sample complexity, thus partially resolving an open problem of Beygelzimer et al. (2015) and Jung et al. (2017).

10.1 Background

Logistic regression is a classical model in statistics used for estimating conditional probabilities (Berkson, 1944). Also known as *conditional maximum entropy model* (Berger et al., 1996), logistic regression has been extensively studied in statistics and machine learning and has been widely used in practice both for binary classification and multi-class classification in a variety of applications.

The basic logistic regression problem consists of learning a linear predictor with performance measured by the *logistic loss*. In the online setting, when the hypothesis class is that of d-dimensional linear predictors with ℓ_2 norm bounded by B, there are two main algorithmic approaches to logistic regression: Online Gradient Descent (Zinkevich, 2003; Shalev-Shwartz and Singer, 2007; Nemirovski et al., 2009), which admits a regret guarantee of $O(B\sqrt{n})$ over n rounds, and Online Newton Step (Hazan et al., 2007), whose regret bound is in $O(de^B \log(n))$. While the latter bound is logarithmic in n, its poor dependence on B makes it weaker and guarantees an improvement only when $B \ll \frac{1}{2}\log(n)$. The question of whether this dependence on B could be improved was posed as an open problem in COLT 2012 by McMahan and Streeter (2012). Hazan et al. (2014) answered this in the negative, showing a lower bound of $\Omega(\sqrt{n})$ for $B \geq \Omega(\log(n))$.

The starting point for this work is a simple observation: In the online setting, the logistic loss, when viewed as a function of the prediction and the true outcome, is 1-mixable (see Section 10.1.1 for definitions). This observation can be used in conjunction with Vovk's Aggregating Algorithm (Vovk, 1995), which leverages mixability in order to achieve regret bounds scaling logarithmically in an appropriate notion of complexity of the space of predictors, and can be implemented in *polynomial time* in relevant parameters using MCMC methods (Section 10.2). Mixability and efficient implementability open the door to fast rates for online logistic regression and related problems via *improper learning*: using predictions that may not be linear in the instances x_ts. This algorithm circumvents the lower bound of Hazan et al. (2014) via improper learning and attains a substantially more favorable regret guarantee of $O(d \log(Bn))$; this is a *doubly-exponential improvement* of the dependence on the scale parameter B. This algorithm provides a positive resolution to to a variant of the open problem of McMahan and Streeter (2012) where improper predictions are allowed.

This improper learning observation leads to a series of new results for agnostic statistical learning. First, we show how to convert the online improper logistic regression algorithm into a agnostic statistical learning algorithm admitting a high-probability excess risk guarantee of $O(d \log(Bn)/n)$ (Section 10.3). While it is straightforward to achieve such a result in expectation using standard online-to-batch conversion techniques, the high-probability bound is more technically challenging. This is achieved using a new technique based on a modified version of the "boosting the confidence" scheme proposed by Mehta (2017) for exp-concave losses. We also prove a lower bound showing that the logarithmic dependence on B of the guarantee of our new algorithm cannot be improved. Finally, we show how to (non-constructively) generalize the $\log(B)$ dependence on predictor norm from linear to arbitrary function classes via sequential symmetrization and chaining arguments (Section 10.7.3). Our general bound indicates that the extent to which dependence on the predictor range B can be improved for general classes is completely determined by their sequential covering numbers.

The basic improper logistic regression algorithm we present in this chapter also resolves two open problems regarding adaptivity to margin in bandit multiclass classification, and boosting. First, the technique provides an algorithm (Section 10.5) for the *online multiclass learning with bandit feedback problem* (Kakade et al., 2008) with an $\tilde{O}(\sqrt{n})$ relative mistake bound with respect to the multiclass logistic loss. This algorithm provides a solution to an open problem of Abernethy and Rakhlin (2009), improving upon the previous algorithm of Hazan and Kale (2011) by providing the $\tilde{O}(\sqrt{n})$ mistake bound guarantee for all possible ranges of parameter sets. Second, the technique provides a new *online multiclass boosting* algorithm (Section 10.7.5) with optimal sample complexity, thus partially resolving an open problem from (Beygelzimer et al., 2015; Jung et al., 2017) (the algorithm is sub-optimal in the number of weak learners it uses, though it is no worse in this regard than previous adaptive algorithms).

10.1.1 Preliminaries

For multiclass classification, the number of output classes is K and the set of output classes is denoted by $[K] := \{1, 2, \ldots, K\}$. Linear predictors are parameterized by weight matrices in $\mathbb{R}^{K\times d}$ so that for an input vector $x \in \mathcal{X}$, $Wx \in \mathbb{R}^K$ is the vector of scores assigned by W to the classes in $[K]$. For a weight matrix W and $k \in [K]$, we denote by W_k the k-th row of W. The space of parameter weight matrices is a convex set $\mathcal{W} \subseteq \{W \in \mathbb{R}^{K\times D} \mid \forall k \in [K], \|W_k\| \leq B\}$ for some known parameter $B > 0$. Thus for all $x \in \mathcal{X}$ and $W \in \mathcal{W}$, we have $\|Wx\|_\infty \leq BR$.

Define the softmax function $\boldsymbol{\sigma} : \mathbb{R}^K \to \Delta_K$ via $\boldsymbol{\sigma}(z)_k = \frac{e^{z_k}}{\sum_{j\in[K]} e^{z_j}}$ for $k \in [K]$. We also define a pseudoinverse for $\boldsymbol{\sigma}$ via $\boldsymbol{\sigma}^+(p)_k = \log(p_k)$ which has the property that for all $p \in \Delta_K$, we have $\boldsymbol{\sigma}(\boldsymbol{\sigma}^+(p)) = p$ and $\sum_{k\in[K]} e^{\boldsymbol{\sigma}^+(p)_k} = 1$. The multiclass logistic loss, also referred to as *softmax-cross-entropy* loss, is defined as $\ell : \mathbb{R}^K \times [K] \to \mathbb{R}$ as $\ell(z, y) := -\log(\boldsymbol{\sigma}(z)_y)$.

It will be convenient to overload notation and define a weighted version of the multiclass logistic loss function as follows: let $\mathcal{Y} := \left\{y \in \mathbb{R}_+^K \mid \|y\|_1 \le L\right\}$ for some known parameter $L > 0$. Then the weighted multiclass logistic loss function $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ is defined by $\ell(z, y) = -\sum_{k\in[K]} y_k \log(\boldsymbol{\sigma}(z)_k)$. It can also be seen by straightforward manipulation that the above definition is equivalent to $\ell(z, y) = \sum_{j\in[K]} y_j \log\Big(1 + \sum_{k\neq j} e^{z_k - z_j}\Big)$.

In the binary classification setting, the standard definition of the logistic loss function is (superficially) different: the label set is is $\{-1, 1\}$, and the logistic loss $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}$ is defined as $\ell_{\text{bin}}(z, y) = \log(1 + \exp(-yz))$. Linear predictors are parameterized by weight vectors $w \in \mathbb{R}^d$ with $\|w\|_2 \le B$, and the loss for a predictor with parameter $w \in \mathbb{R}^d$ on an example $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ is $\ell_{\text{bin}}(\langle w, x\rangle, y)$. This loss can be equivalently viewed in the multiclass framework above setting $K = 2$, $\mathcal{W} = \{W \in \mathbb{R}^{2\times d} \mid \|W_1\|_2 \le B, W_2 = 0\}$, and mapping the labels $1 \mapsto 1$ and $-1 \mapsto 2$.

Finally, we make frequent use of a smoothing operator $\mathrm{smooth}_\mu : \Delta_K \to \Delta_K$ for a parameter $\mu \in [0, 1/2]$, defined via $\mathrm{smooth}_\mu(p) = (1 - \mu)p + \mu\mathbf{1}/K$ where $\mathbf{1} \in \mathbb{R}^K$ is the all ones vector. We use the notation $\mathbb{1}[\cdot]$ to denote the indicator random variable for an event.

Online Multiclass Logistic Regression. We use the following multiclass logistic regression protocol. Learning proceeds over a series of rounds indexed by $t = 1, \ldots, n$. In each round t, nature provides $x_t \in \mathcal{X}$, and the learner selects prediction $\hat{z}_t \in \mathbb{R}^K$ in response.[1] Then nature provides an outcome $y_t \in [K]$ or $y_t \in \mathcal{Y}$, depending on application, and the learner incurs multiclass logistic loss $\ell(\hat{z}_t, y_t)$. The regret of the learner is defined to be

[1]We use the notation $\hat{z}_t$ for predictions made in the logistic regression setting to keep $\widehat{y}_t$ reserved for downstream applications of the logistic regression algorithm.

$\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{W \in \mathcal{W}} \sum_{t=1}^{n} \ell(W x_t, y_t)$.

The learner is said to be *proper* if it generates $\hat{z}_t$ by choosing a weight matrix $W_t \in \mathcal{W}$ *before* observing the pair (x_t, y_t) and setting $\hat{z}_t = W_t x_t$. This is the standard protocol when the problem is viewed as an instance of online convex optimization, and is the setting for previous investigations into fast rates for logistic regression (Bach, 2010; McMahan and Streeter, 2012; Bach and Moulines, 2013; Bach, 2014), including the negative result of Hazan et al. (2014). The more general online learning setting that is described above allows *improper* learners which may generate $\hat{z}_t$ arbitrarily using knowledge of x_t.

Fast Rates and Mixability. Conditions under which *fast rates* for online/statistical learning (meaning that average regret or generalization error scales as $\tilde{O}(1/n)$ rather than $O(1/\sqrt{n})$) are achievable have been studied extensively (see (Van Erven et al., 2015) and the references therein). For the purpose of this chapter, a rather general condition on the structure of the problem that leads to fast rates is Vovk's notion of *mixability* (Vovk, 1995), which we define in an abstract setting below. Consider a prediction problem where the set of outcomes is $\mathcal{Y}$ and the set of predictions is $\mathcal{Z}$, and the loss of a prediction on an outcome is given by a function $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$. For a parameter $\eta > 0$, the loss function ℓ is said to be η-mixable if for any probability distribution π over $\mathcal{Z}$, there exists a *"mixed"* prediction $z_\pi \in \mathcal{Z}$ such that for all possible outcomes $y \in \mathcal{Y}$, we have $\mathbb{E}_{z \sim \pi}[\exp(-\eta \ell(z, y))] \leq \exp(-\eta \ell(z_{\text{mix}}, y))$.

Now suppose that we are given a finite reference class of predictors $\mathcal{F}$ consisting of functions $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{X}$ is the input space. The problem of online learning over $\mathcal{F}$ with an η-mixable loss function admits an *improper* algorithm, viz. Vovk's Aggregating Algorithm (Vovk, 1995), with regret bounded by $\frac{\log |\mathcal{F}|}{\eta}$, a *constant* independent of the number of prediction rounds n. The algorithm simply runs the standard exponential weights/Hedge algorithm (Cesa-Bianchi and Lugosi, 2006) with learning rate set to η. In each round t, given an input x_t, the distribution over $\mathcal{F}$ generated by the exponential weights algorithm induces a distribution over $\mathcal{Z}$ via the outputs of the predictors on x_t, and the Aggregating Algorithm plays the mixed prediction for this distribution over $\mathcal{Z}$. Finally, if $\mathcal{F}$ is infinite, under appropriate conditions on $\mathcal{F}$ fast rates can be obtained by running a continuous version of the same algorithm. This is the strategy we employ in this chapter for the logistic loss.

10.2 Improved Rates for Online Logistic Regression

We start by providing a simple proof of the mixability of the multiclass logisitic loss function for the case when the outcomes y is a class in $[K]$ (i.e. the unweighted case).

Proposition 15. The unweighted multiclass logistic loss $\ell : \mathbb{R}^K \times [K] \to \mathbb{R}$ defined as $\ell(z, y) = -\log(\boldsymbol{\sigma}(z)_y)$ is 1-mixable.

Proof. The proof is by construction. Given a distribution π on $\mathbb{R}^K$, define $z_\pi = \boldsymbol{\sigma}^+(\mathbb{E}_{z\sim\pi}[\boldsymbol{\sigma}(z)])$. Now, for any $y \in [K]$, we have $\mathbb{E}_{z\sim\pi}[\exp(-\ell(z, y))] = \mathbb{E}_{z\sim\pi}[\boldsymbol{\sigma}(z)_y] = \boldsymbol{\sigma}(z_\pi)_y = \exp(-\ell(z_\pi, y))$. The second equality above uses the fact that for any $p \in \Delta_K$, $\boldsymbol{\sigma}(\boldsymbol{\sigma}^+(p)) = p$. Thus, ℓ is 1-mixable. $\square$

With a little more work, we can prove that the weighted multiclass logistic loss function is also mixable with a constant that inversely depends on the total weight. The proof appears in Section 10.7.

Proposition 16. Let $\mathcal{Y} := \left\{y \in \mathbb{R}_+^K \mid \|y\|_1 \le L\right\}$ for some parameter $L > 0$. The weighted multiclass logistic loss $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ defined as $\ell(z, y) = -\sum_{k\in[K]} y_k \log(\boldsymbol{\sigma}(z)_k)$ is $\frac{1}{L}$-

mixable. For any distribution π on $\mathbb{R}^K$, the mixed prediction $z_\pi = \boldsymbol{\sigma}^+(\mathbb{E}_{z\sim\pi}[\boldsymbol{\sigma}(z)])$ certifies $\frac{1}{L}$-mixability of ℓ.

We are now ready to state a variant of Vovk's Aggregating Algorithm, Algorithm 9 for the online multiclass logistic regression problem from Section 10.1.1, operating over a class of linear predictors parameterized by weight matrices W in some convex set $\mathcal{W}$. The algorithm and its regret bound (proved in Section 10.7) are given in fairly general form that is useful for subsequent applications.

Algorithm 9

```
procedure (decision set 𝒲, smoothing parameter μ ∈ [0, 1/2].)
    Initialize P_1 to be the uniform distribution over 𝒲.
    for t = 1, ..., n do
        Obtain x_t and predict ẑ_t = σ⁺(smooth_μ(E_{W~P_t}[σ(W x_t)])).
        Obtain y_t and define P_{t+1} as the distribution over 𝒲 with density
            P_{t+1}(W) ∝ exp(−(1/L) Σ_{s=1}^{t} ℓ(W x_s, y_s)).
    end for
end procedure
```

Theorem 32. *The regret of Algorithm 9 is bounded by*

$$\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{W\in\mathcal{W}} \sum_{t=1}^{n} \ell(Wx_t, y_t) \leq 5LD_{\mathcal{W}} \cdot \log\left(\frac{BRn}{D_{\mathcal{W}}} + e\right) + 2\mu \sum_{t=1}^{n} \|y_t\|_1, \tag{10.1}$$

where $D_{\mathcal{W}} := \dim(\mathcal{W}) \leq dK$ is the linear-algebraic dimension of $\mathcal{W}$. The predictions $(\hat{z}_t)_{t\leq n}$ generated by the algorithm satisfy $\|\hat{z}_t\|_\infty \leq \log(K/\mu)$.

Increasing the smoothing parameter μ only degrades the performance of Algorithm 9. However, smoothing ensures that each prediction $\hat{z}_t$ is bounded, which is important for our applications.

For the special case of multiclass prediction when $y \in [K]$, this algorithm enjoys a regret bound of $O(dK \log(\frac{BRn}{dK} + e))$. It thus provides a positive resolution to the open problem of McMahan and Streeter (2012) (in fact, with an exponentially better dependence on B than what the open problem asked for), using improper predictions to circumvent the lower bound of Hazan et al. (2014).

Turning to efficient implementation, it has been noted (e.g. (Hazan et al., 2007)) that log-concave sampling or integration techniques (Lovász and Vempala, 2006, 2007) can be applied to compute the expectation in Algorithm 9 in polynomial time. The following proposition makes this idea rigorous[2] and is proven formally in Section 10.7.6. We note that this is not a practical algorithm, however, and obtaining a truly practical algorithm with a modest polynomial dependence on the dimension is a significant open problem.

Proposition 17. Algorithm 9 can be implemented approximately so that the regret bound (10.1) is obtained up to additive constants in time $\mathrm{poly}(d, n, B, R, K, L)$.

Finally, to conclude this section we state a lower bound, which shows that the $\log(B)$ factor in the regret bound in Theorem 32 cannot be improved for most values of B. This lower bound is by reduction to learning halfspaces with a margin in a Perceptron-type setting: We first show that Algorithm 9 can be configured to give a mistake bound of $O(d\log(\log(n)/\gamma))$ for binary classification with halfspaces and margin γ,[3] then give a lower bound against this type of rate.

For simplicity, the lower bound is only stated in the binary outcome settting and we use the standard definition of the binary logistic loss, ℓ_{bin} from Section 10.1.1. The proof is in Section 10.7.

Theorem 33 (Lower bound)**.** *Consider the binary logistic regression problem over the class of linear predictors with parameter set* $\mathcal{W} = \{w \in \mathbb{R}^d \mid \|w\|_2 \leq B\}$ *with* $B = \Omega(\sqrt{d}\log(n))$*. Then for any algorithm for prediction with the binary logistic loss, there is a sequence of examples*

[2] A subtlety is that since $\hat{z}_t$ is evaluated inside the nonlinear logistic loss we cannot exploit linearity of expectation.

[3] It is a folklore result that this type of margin bound can be obtained by running a variant of the ellipsoid method online.

$(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$ *for* $t \in [n]$ *with* $\|x_t\|_2 \leq 1$ *such that the regret of the algorithm is* $\Omega\left(d \log\left(\frac{B}{\sqrt{d}\log(n)}\right)\right)$.

Relation to Bayesian model averaging To the best of our knowledge, the mixability of the logistic loss has surprisingly not appeared in the literature. However, Algorithm 9 can be seen as an instance of Bayesian model averaging, and consequently the analysis of Kakade and Ng (2005) can be applied to derive the same $O(d \log(Bn/d))$ regret bound as in Theorem 32 in the binary setting. Specifically, it suffices to apply their Theorem 2.2 with parameter $\nu^2 = B^2/d$. This highlights that Bayesian approaches can have great utility even when analyzed outside of the Bayesian framework.

10.3 Agnostic Statistical Learning Guarantees

Before the results in this chapter were developed, the issue of improving on the $O(e^B)$ fast rate for logistic regression was not addressed even in the i.i.d. statistical learning setting (Section 2.3). This is perhaps not surprising since the proper lower bound proven by Hazan et al. (2014) applies in this setting as well.

Using our improved online algorithm as a starting point, we will show that it is possible to obtain a predictor with excess risk bounded in *high-probability* by $O(d \log(Bn)/n)$ for the batch logistic regression problem. While it is quite straightforward to show that the standard online-to-batch conversion technique applied to Algorithm 9 provides a predictor that obtains such an excess risk bound in expectation, obtaining a high-probability bound is far less trivial, as we must ensure that deviations scale at most as $O(\log(B))$. Indeed, a different algorithm is necessary, and our approach is to use a modified version of the "boosting the confidence" scheme proposed by Mehta (2017) for exp-concave losses. Our main result for linear classes is Theorem 34 below. For notational convenience will use the shorthand $\mathbb{E}_{(x,y)}[\cdot]$ to denote $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\cdot]$ where $\mathcal{D}$ is an unknown distribution over $\mathcal{X} \times [K]$.

Theorem 34 (High-probability excess risk bound)**.** *Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X} \times [K]$. For any $\delta > 0$ and n samples $\{(x_t, y_t)\}_{t=1}^n$ drawn from $\mathcal{D}$, we can construct $g : \mathcal{X} \to \mathbb{R}^K$ such that w.p. at least $1-\delta$, the excess risk $\mathbb{E}_{(x,y)}[\ell(g(x), y)] - \inf_{W\in\mathcal{W}} \mathbb{E}_{(x,y)}[\ell(Wx, y)]$ is bounded by*

$$O\left(\frac{dK \log\left(\frac{BRn}{\log(1/\delta)dK} + e\right)\log\left(\frac{1}{\delta}\right) + \log(Kn)\log\left(\frac{\log(n)}{\delta}\right)}{n}\right).$$

Theorem 34 is a consequence of the more general Theorem 39—stated and proved in Section 10.7.2—concerning prediction with the log loss $\ell_{\log} : \Delta_K \times [K] \to \mathbb{R}$ defined as $\ell_{\log}(p, y) = -\log(p_y)$. The theorem asserts that we can convert any online algorithm for multiclass learning with log loss that predicts distributions in Δ_K for any given input into a predictor for the batch problem with an excess bound essentially equal to the average regret with high probability.

10.4 Minimax Bounds for General Function Classes

We now turn to the question of extending our techniques to general, non-linear predictors. We characterize the minimax regret for learning with the unweighted multiclass logistic loss[4] for a general class $\mathcal{F}$ of predictors $f : \mathcal{X} \to \mathbb{R}^K$ and abstract instance space $\mathcal{X}$. This is the same setting as in Section 10.1.1, but with the benchmark class $\{x \mapsto Wx \mid W \in \mathcal{W}\}$ replaced with an arbitrary class $\mathcal{F}$, where the loss of a predictor $f \in \mathcal{F}$ on an example $(x, y) \in \mathcal{X} \times [K]$ is given by $\ell(f(x), y) = -\log(\boldsymbol{\sigma}(f(x))_y)$. The bounds we present in this

[4]We only consider the unweighted case in this section to avoid excessive notation.

section—based on sequential covering numbers—are based on bounding a martingale process arising via minimax analysis, as in Part II.

Specializing the setup of Section 2.3 to the logistic loss, the minimax regret is written as

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}) = \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{\hat{z}_t \in \mathbb{R}^K} \max_{y_t \in [K]} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right]. \tag{10.2}$$

Our bounds on $\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F})$ exploit that the logistic loss can be viewed in two complementary ways: since the loss is 1-mixable, one can attain a bound of $O(\log|\mathcal{F}|)$ for finite function classes $\mathcal{F}$ using the Aggregating Algorithm, and since the loss is 2-Lipschitz (in the ℓ_∞ norm), for more complex classes one can obtain bounds using sequential Rademacher complexity (Rakhlin et al., 2014). Our analysis uses both properties simultaneously.

Here is a sketch of the idea for a special case in which we make the simplifying assumption that $\mathcal{F}$ admits a pointwise cover. Recall that a pointwise cover for $\mathcal{F}$ at scale γ is a set V of functions $g : \mathcal{X} \to \mathbb{R}^K$ such that for any $f \in \mathcal{F}$, there is a $g \in V$ such that for all $x \in \mathcal{X}$, $\|f(x) - g(x)\|_\infty \leq \gamma$. Let $N(\gamma)$ be the size of a minimal such cover. For every $g \in V$, let $\mathcal{F}_g = \{f \in \mathcal{F} \mid \sup_{x \in \mathcal{X}} \|f(x) - g(x)\|_\infty \leq \gamma\}$. Now consider the following two-level algorithm. Within each $\mathcal{F}_g$, run the minimax online learning algorithm for this set, then aggregate the predictions for these algorithms over all $g \in V$ using the Aggregating Algorithm to produce the final prediction $\hat{z}_t$.

For each $g \in V$, the regret of the minimax optimal online learning algorithm competing with $\mathcal{F}_g$ can be bounded by the sequential Rademacher complexity of $\mathcal{F}_g$, which can in turn be bounded by the Dudley integral complexity using that the loss is 2-Lipschitz and that the L_∞ "radius" of $\mathcal{F}_g$ is at most γ (Rakhlin et al., 2014). The Aggregating Algorithm, via 1-mixability, ensures a regret bound of $\log N(\gamma)$ against any sub-algorithm. This algorithm has the following regret bound:

$$\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \le \inf_{\gamma>0} \left\{ \log N(\gamma) + \inf_{\alpha>0} \left\{ 8\alpha n + 24\sqrt{n} \int_{\alpha}^{\gamma} \sqrt{\log N(\delta)} d\delta \right\} \right\}. \tag{10.3}$$

This procedure already yields the same bound for the d-dimensional linear setting explored earlier: For a class $x \mapsto Wx$ with $\|W\|_2 \le B$ it holds that $N(\gamma) \le \left(\frac{B}{\gamma}\right)^{Kd}$, and we can use this bound in conjunction with (10.3) and the setting $\alpha = \gamma = 1/n$ to get the desired regret bound of $O(dK \log(Bn/dK))$ on the minimax regret.

Unfortunately, this simple approach fails on classes $\mathcal{F}$ for which the pointwise cover is infinite. This can happen for well-behaved function classes that have small *sequential covering number*, even though bounded sequential covering number is sufficient for learnability in the online setting (Rakhlin et al., 2014). We now provide a bound that replaces the pointwise covering number in the argument above with the sequential covering number. Since we work in the multiclass setting we require a slight generalization of the sequential covering number definition from Chapter 6.

Definition 9. *For any set $\mathcal{Z}$, a $\mathcal{Z}$-valued K-ary tree of depth n is a sequence $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ of mappings with $\mathbf{z}_t : [K]^{t-1} \to \mathcal{Z}$.*

Definition 10. *A set V of $\mathbb{R}^K$-valued K-ary trees is an α-cover (w.r.t. the L_p norm) of $\mathcal{F}$ on an $\mathcal{X}$-valued K-ary tree $\mathbf{x}$ of depth n with loss ℓ if*

$$\forall f \in \mathcal{F},\ y \in [K]^n,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{n} \sum_{t=1}^{n} \max_{y'_t \in [K]} |\ell(f(\mathbf{x}_t(y)), y'_t) - \ell(\mathbf{v}_t(y), y'_t)|^p \right)^{1/p} \le \alpha.$$

Definition 11. *The L_p covering number of $\mathcal{F}$ on tree $\mathbf{x}$ is defined as*

$$\mathcal{N}_p(\alpha, \ell \circ \mathcal{F}, \mathbf{x}) := \min\{|V| : V \textit{ is an } \alpha\text{-cover of } \mathcal{F} \text{ on } \mathbf{x} \text{ w.r.t. the } L_p \text{ norm}\}.$$

Further, define $\mathcal{N}_p(\alpha, \ell \circ \mathcal{F}) = \sup_{\mathbf{x}} \mathcal{N}_p(\alpha, \ell \circ \mathcal{F}, \mathbf{x})$.

If $K = 2$ then the above definition coincides with the previous sequential cover definition for real valued function classes.

Theorem 35. *Any function class $\mathcal{F}$ that is uniformly bounded*[5] *over $\mathcal{X}$ enjoys the minimax value bound:*

$$\mathcal{V}_n^{\text{ol}}(\mathcal{F}) \leq \inf_{\gamma>0} \left\{ \log \mathcal{N}_2(\gamma, \ell \circ \mathcal{F}) + \inf_{\gamma \geq \alpha > 0} \left\{ 8\alpha n + 24\sqrt{n} \int_{\alpha}^{\gamma} \sqrt{\log(\mathcal{N}_2(\delta, \ell \circ \mathcal{F}) \cdot n)} d\delta \right\} \right\} + 4. \tag{10.4}$$

This rate overcomes several shortcomings faced when trying to apply previously developed minimax bounds for general function classes to the logistic loss. Specifically, Rakhlin et al. (2014) applies to our logistic loss setup but ignores the curvature of the loss and so cannot obtain fast rates, while Rakhlin and Sridharan (2015) obtain fast rates but scale with e^B, where B is a bound on the magnitude of the predictions, because they use exp-concavity.

Our general function class bound is especially interesting in light of rates obtained in Rakhlin and Sridharan (2014) for the square loss, which are also based on sequential covering numbers. In the binary case the bound (10.4) precisely matches the general class bound of (Rakhlin and Sridharan, 2014, Lemma 5) in terms of dependence on the sequential metric entropy. However, (10.4) does not depend on B explicitly, whereas their Lemma 5 bound for the square loss explicitly scales with B^2. In other words, compared to other common curved losses the logistic loss has a desirable property:

The minimax rate for logistic regression only depends on scale through capacity of the class $\mathcal{F}$.

Let us examine some rates obtained from this bound for concrete settings. These examples are based on sequential covering bounds that appeared in Rakhlin and Sridharan (2014, 2015).

[5]Boundedness is used to apply the minimax theorem, but does not explicitly enter our quantitative bounds.

Example 21 (Sparse linear predictors). *Let $\mathcal{G} = \{g_1, \ldots, g_M\}$ be a set of M functions $g_i : \mathcal{X} \mapsto [-B, B]$. Define $\mathcal{F}$ to be the set of all convex combinations of at most s out of these M functions. The sequential covering number can be easily upper bounded: We can choose s out of M functions in $\binom{M}{s}$ ways. For each choice, the sequential covering number for the set of all convex combinations of these s bounded functions at scale β is bounded as $\frac{B^s}{\beta^s}$. Hence, using that the logistic loss is Lipschitz, we conclude that $\mathcal{N}_2(\mathcal{F}, \beta) = O\big(\big(\frac{eM}{s}\big)^s \cdot \beta^{-s} B^s\big)$. Using this bound with Theorem 35 we obtain $\mathcal{V}_n(\mathcal{F}) \leq O\left(s \log(BMn/s)\right)$.*

The bounds from Rakhlin et al. (2014); Rakhlin and Sridharan (2014, 2015) either pay $O(B\sqrt{n})$ or $O(e^B)$ on this example, whereas the new bound from (10.4) correctly obtains $O(\log(B))$ scaling.

Example 22 (Besov classes). *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Let $\mathcal{F}$ be the ball of radius B in Besov space $B^s_{p,q}(\mathcal{X})$. When $s > d/p$ it can be shown that the pointwise log covering number of the space at scale β is of order $(B/\beta)^{d/s}$. When $p \geq 2$ one can obtain a sequential covering number bound of order $(B/\beta)^p$ (Rakhlin and Sridharan, 2015, Section 5.8). These bounds imply:*

1. *If $s \geq d/2$, then $\mathcal{V}_n(\mathcal{F}) \leq \tilde{O}\left(B^{\frac{2d}{d+2s}} n^{\frac{d}{d+2s}}\right)$.*
2. *$s < d/2$, then: if $p > 1 + d/2s$ then $\mathcal{V}_n(\mathcal{F}) \leq \tilde{O}\left(Bn^{1-\frac{s}{d}}\right)$; if not, $\mathcal{V}_n(\mathcal{F}) \leq \tilde{O}(Bn^{1-1/p})$.*

Remark 2. *Using the machinery from the previous section, we can generically lift the general function class bounds given by Theorem 35 to high-probability bounds for the i.i.d. batch setting.*

10.5 Application: Bandit Multiclass Learning

We now apply the logistic regression machinery we have developed to the *bandit multiclass classification* problem. This problem, first studied by Kakade et al. (2008), considers the protocol of online multiclass learning in Section 10.1.1 with nature choosing $y_t \in [K]$ in each round, but with the added twist of bandit feedback: in each round, the learner predicts a class $\widehat{y}_t \sim p_t$ and receives feedback only on whether the prediction was correct or not, i.e. $\mathbb{1}[\widehat{y}_t \neq y_t]$. The goal is to minimize regret with respect to a reference class of linear predictors, using some appropriate surrogate loss function for the 0-1 loss.

Kakade et al. (2009b) used the multiclass hinge loss $\ell_{\text{hinge}}(W, (x_t, y_t)) = \max_{k \in [K] \setminus \{y_t\}}[1 + \langle W_k, x_t \rangle - \langle W_{y_t}, x_t \rangle]_+$ and gave an algorithm based on the multiclass Perceptron algorithm achieving $O(n^{2/3})$ regret. For a Lipschitz continuous surrogate loss function, running the EXP4 algorithm (Auer et al., 2002b) on a suitable discretization of the space of all linear predictors obtains $\tilde{O}(\sqrt{n})$ regret, albeit very inefficiently, i.e. with exponential dependence on the dimension. In COLT 2009, Abernethy and Rakhlin (2009) posed the open problem of obtaining an *efficient* algorithm for the problem with $O(\sqrt{n})$ regret. Specifically, they suggested the multiclass logistic loss as an appropriate surrogate loss function for the problem. Hazan and Kale (2011) solved the open problem and obtained an algorithm, Newtron, based on the Online Newton Step algorithm (Hazan et al., 2007) with $\tilde{O}(\sqrt{n})$ regret for the case when norm of the linear predictors scales at most logarithmically in n. Beygelzimer et al. (2017) also solved the open problem presenting a different algorithm called SOBA. SOBA is analyzed using a different family of surrogate loss functions parameterized by a scalar $\eta \in [0, 1]$ with $\eta = 0$ corresponding to the hinge loss and $\eta = 1$ corresponding to the squared hinge loss. For all values of $\eta \in [0, 1]$, SOBA simultaneously obtains relative bound mistake bounds of $\tilde{O}(\frac{1}{\eta}\sqrt{n})$ with the comparator's loss measured with respect to the corresponding loss function.

Now we present an algorithm, OBAMA (for *Online Bandit Aggregation Multiclass Algorithm*), depicted in Algorithm 10 in Section 10.7.4, that obtains an $\tilde{O}(\sqrt{n})$ relative mistake bound for the multiclass logistic loss, thus providing another solution to the open problem of

Algorithm 10

procedure OBAMA(decision set $\mathcal{W}$, smoothing parameter μ.)
  Let $\mathcal{A}$ be Algorithm 9 initialized with $\mathcal{W}$ and μ.
  for $t = 1, \ldots, n$ **do**
    Obtain x_t, pass it to $\mathcal{A}$ and let $\hat{z}_t \in \mathbb{R}_K$ be the output of $\mathcal{A}$.
    Play $\widehat{y}_t \sim p_t := \boldsymbol{\sigma}(\hat{z}_t)$ and obtain $\mathbb{1}[\widehat{y}_t \neq y_t]$.
    Define $\tilde{y}_t \in \mathbb{R}^K$ as $\tilde{y}_t(k) := \frac{\mathbb{1}[k=\widehat{y}_t]\mathbb{1}[\widehat{y}_t=y_t]}{p_t(\widehat{y}_t)}$ for $k \in [K]$ and pass it as feedback to $\mathcal{A}$.
  end for
end procedure

Abernethy and Rakhlin (2009). The mistake bound of OBAMA trumps that of Newtron, since both algorithms rely on the same loss function, and OBAMA obtains an $\tilde{O}(\sqrt{n})$ relative mistake bound on a larger range of parameter values compared to Newtron. While SOBA also has an $\tilde{O}(\sqrt{n})$ relative mistake bound, the two bounds are incomparable since they are relative to the comparator's loss measured using different loss functions.

Theorem 36. *There is a setting of the smoothing parameter μ such that OBAMA enjoys the following mistake bound:*

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{1}[\widehat{y}_t \neq y_t] \\
&\leq \inf_{W \in \mathcal{W}} \sum_{t=1}^{n} \ell(Wx_t, y_t) + O\left(\min\left\{dK^2 e^{2BR} \log\left(\tfrac{BRn}{dK} + e\right),\ \sqrt{dK^2 \log(\tfrac{BRn}{dK} + e)n}\right\}\right).
\end{aligned}$$

This bound significantly improves upon that of Newtron (Hazan and Kale, 2011), which is of order $O(dK^3 \min\{\exp(BR)\log(n), BRn^{\frac{2}{3}}\})$ under the same setting and surrogate loss. The proof of Theorem 36 appears in Section 10.7.4.

10.6 Application: Online Multiclass Boosting

The final application of our online logistic regression results is to derive adaptive online boosting algorithms with optimal sample complexity, which improves the AdaBoost.OL algorithm of Beygelzimer et al. (2015) for the binary classification setting as well as its multiclass extension AdaBoost.OLM of Jung et al. (2017). We state our improved online boosting algorithm in the multiclass setting for maximum generality, following the exposition and notation of Jung et al. (2017) fairly closely.

We consider the following online multiclass prediction setting with 0-1 loss. In each round t, for $t = 1, \ldots, n$, the learner receives an instance $x_t \in \mathcal{X}$, then selects a class $\widehat{y}_t \in [K]$, and finally observes the true class $y_t \in [K]$. The goal is to minimize the total number of mistakes $\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\}$.

In the boosting setup, we are interested in obtaining strong mistake bounds with the help of *weak learners*. Specifically, the learner is given access to N copies of a weak learning algorithm for a cost-sensitive classification task. Each weak learner $i \in [N]$ works in the following protocol: for time $t = 1, \ldots, n$, 1) receive $x_t \in \mathcal{X}$ and cost matrix $C_t^i \in \mathcal{C}$; 2) predict class $l_t^i \in [K]$; 3) receive true class $y_t \in [K]$ and suffer loss $C_t^i(y_t, l_t^i)$. Here $\mathcal{C}$ is some fixed cost matrices class and we follow (Jung et al., 2017) to restrict to $\mathcal{C} = \left\{ C \in \mathbb{R}_+^{K \times K} \mid \forall y \in [K], C(y,y) = 0 \text{ and } \|C(y, \cdot)\|_1 \leq 1 \right\}$.

To state the weak learning condition, we define a randomized baseline $u_{\gamma,y} \in \Delta_K$ for some edge parameter $\gamma \in [0,1]$ and some class $y \in [K]$, so that $u_{\gamma,y}(k) = (1-\gamma)/K$ for $k \neq y$ and $u_{\gamma,y}(k) = (1-\gamma)/K + \gamma$ for $k = y$. In other words, $u_{\gamma,y}$ puts equal weight to all classes except for the class y which gets γ more weight. The assumption we impose on the weak learners is then that their performance is comparable to that of a baseline which always picks the true class with slightly higher probability than the others, formally stated below.

Definition 12 (Weak Learning Condition (Jung et al., 2017))**.** *An environment and a learner outputting $(l_t)_{t \leq n}$ satisfy the multiclass weak learning condition with edge γ and sample complexity S if for all outcomes $(y_t)_{t \leq n}$ and cost matrices $(C_t)_{t \leq n}$ from the set $\mathcal{C}$ adaptively chosen by the* *environment, we have* $\sum_{t=1}^{n} C_t(y_t, l_t) \leq \sum_{t=1}^{n} \mathbb{E}_{k\sim u_{\gamma,y_t}}[C_t(y_t, k)] + S$.[6]

10.6.1 AdaBoost.OLM++

The high-level idea of our algorithm is similar to that of AdaBoost.OL and AdaBoost.OLM: find a weighted combination of weak learners to minimize some version of the logistic loss in an online manner. The key difference is that previous works use simple gradient descent to find the weight for each weak learner via proper learning, while we translate the problem into the framework discussed in Section 10.2 and deploy the proposed improper learning techniques to obtain an improvement on the regret for learning these weights, which then leads to better and in fact optimal sample complexity.

Another difference compared to (Jung et al., 2017) is that the logistic loss we use here is more suitable for the multiclass problem than the one they use.[7] This simple modification leads to exponential improvement in the number of classes K for the number of weak learners required.

We now describe our algorithm, called AdaBoost.OLM++, in more detail (see Algorithm 11 in Section 10.7.5). We denote the i-th weak learner as WL^i, which is seen as a stateful object and supports two operations: WL^i.Predict(x, C) predicts a class given an instance and a cost matrix but does not update its internal state; WL^i.Update(x, C, y) updates the state given an instance, a cost matrix and the true class y. To keep track of the state we use the notation WL^i_t to imply that it has been updated for $t-1$ times.

For each weak learner, the algorithm also maintains an instance of Algorithm 9, denoted by

[6] This is in fact a weaker weak learning condition than that of (Jung et al., 2017), which also allows weights.

[7] The loss Jung et al. (2017) use moves the sum over the incorrect classes outside the log, that is, $\ell(z, y) = \sum_{k\neq y} \log(1 + e^{z_k - z_y})$.

Logistic^i, to improperly learn the aforementioned weight for this weak learner. Similarly, we use Logistic^i.Predict(x) to denote the prediction step (step 4) in Algorithm 9 and Logistic^i.Update(x, y) to denote the update step (i.e. step 5). The notation Logistic^i_t again implies that the state has been updated $t-1$ times.

Algorithm 11 AdaBoost.OLM++

procedure AdaBoost.OLM++(weak learners $\text{WL}^1, \ldots, \text{WL}^N$)
  For all $i \in [N]$, set $v_1^i \leftarrow 1$, initialize weak learner WL_1^i, and initialize logistic learner Logistic$_1^i$ with $\mathcal{W} = \left\{(\alpha I_{K\times K}, I_{K\times K}) \in \mathbb{R}^{K\times 2K} \mid \alpha \in [-2, 2]\right\}$ and $\mu = 1/n$.
  for $t = 1, \ldots, n$ **do**
    Receive instance x_t.
    $s_t^0 \leftarrow 0 \in \mathbb{R}^K$.
    for $i = 1, \ldots, N$ **do**
      Compute cost matrix C_t^i from s_t^{i-1} using (10.5).
      $l_t^i \leftarrow \text{WL}_t^i.\text{Predict}(x_t, C_t^i)$.
      $\widetilde{x}_t^i \leftarrow (e_{l_t^i}, s_t^{i-1}) \in \mathbb{R}^{2K}$.
      $s_t^i \leftarrow \text{Logistic}_t^i.\text{Predict}(\widetilde{x}_t^i)$.
      $\widehat{y}_t^i \leftarrow \arg\max_k s_t^i(k)$.
    end for
    Sample i_t with $\mathbb{P}(i_t = i) \propto v_t^i$.
    Predict $\widehat{y}_t = \widehat{y}_t^{i_t}$ and receive true class $y_t \in [K]$.
    for $i = 1, \ldots, N$ **do**
      $\text{WL}_{t+1}^i \leftarrow \text{WL}_t^i.\text{Update}(x_t, C_t^i, y_t)$.
      $\text{Logistic}_{t+1}^i \leftarrow \text{Logistic}_t^i.\text{Update}(\widetilde{x}_t^i, \mathbb{1}_{y_t})$.
      $v_{t+1}^i \leftarrow v_t^i \cdot \exp(-\mathbb{1}\{\widehat{y}_t^i \neq y_t\})$.
    end for
  end for
end procedure

Our algorithm maintains a variable $s_t^i \in \mathbb{R}^K$ which stands for the weighted accumulated scores of the first i weak learners for instance x_t. When updating s_t^i from s_t^{i-1} given the prediction $l_t^i \in [K]$ of weak learner i, our goal is to have the total loss $\sum_{t=1}^n \ell(s_t^i, y_t)$ close to $\sum_{t=1}^n \ell(s_t^{i-1} + \alpha e_{l_t^i}, y_t)$ for the best α within some range ($[-2, 2]$ suffices). Previous works therefore try to learn this weight α via standard online learning approaches. However, realizing $s_t^{i-1} + \alpha e_{l_t^i}$ can be written as $W\widetilde{x}_t^i$ for $W = (\alpha I_{K\times K}, I_{K\times K}) \in \mathbb{R}^{K\times 2K}$ and $\widetilde{x}_t^i = (e_{l_t^i}, s_t^{i-1}) \in \mathbb{R}^{2K}$, in light of Theorem 32 we can in fact apply Algorithm 9 to learn s_t^i if we let the decision set be $\mathcal{W} = \left\{(\alpha I_{K\times K}, I_{K\times K}) \in \mathbb{R}^{K\times 2K} \mid \alpha \in [-2,2]\right\}$. To make sure that $\widetilde{x}_t^i$ has bounded norm, we also set the smoothing parameter μ to be $1/n$.

With the weighted score s_t^i, the prediction coming from the first i weak learner is naturally define as $\widehat{y}_t^i = \arg\max_k s_t^i(k)$, the class with the largest score. As in AdaBoost.OL and AdaBoost.OLM, these predictions $(\widehat{y}_t^i)_{i\leq N}$ are treated as N experts and the final prediction y_t is determined by the classic Hedge algorithm (Freund and Schapire, 1997) over these experts (Lines 13 and 18).

Finally, the cost matrices fed to the weak learners are closely related to the gradient of the loss function. Formally, define the auxiliary cost matrix $\widehat{C}_t^i$ such that $\widehat{C}_t^i(y,k) = \frac{\partial \ell(z,y)}{\partial z_k}|_{z=s_t^{i-1}}$, which is simply $\boldsymbol{\sigma}(s_t^{i-1})_k$ for $k \neq y$ and $\boldsymbol{\sigma}(s_t^{i-1})_y - 1$ otherwise. The actual cost matrix is then a translated and scaled version of $\widehat{C}_t^i(y,k)$ so that it belongs to the class $\mathcal{C}$:

$$C_t^i(y,k) = \frac{1}{K}\left(\widehat{C}_t^i(y,k) - \widehat{C}_t^i(y,y)\right) \in \mathcal{C}. \tag{10.5}$$

We now give a mistake bound for AdaBoost.OLM++, which holds even without the weak learning condition and is adaptive to the empirical edge of the weak learners.[8] All proofs in this section appear in Section 10.7.5.

Theorem 37. *With probability at least $1-\delta$, the predictions $(\widehat{y}_t)_{t\leq n}$ generated by Algorithm 11 satisfy*

$$\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} = \tilde{O}\left(\frac{n}{\sum_{i=1}^N \gamma_i^2} + \frac{N}{\sum_{i=1}^N \gamma_i^2}\right), \tag{10.6}$$

where $\gamma_i = \frac{\sum_{t=1}^n \widehat{C}_t^i(y_t, l_t^i)}{\sum_{t=1}^n \widehat{C}_t^i(y_t, y_t)} \in [-1,1]$ is the empirical edge of weak learner i.

We can now relate the empirical edges to the edge defined in the weak learning condition.

[8] In this chapter we use the notation $\tilde{O}$ and $\tilde{\Omega}$ to hide dependence logarithmic in n, N, K and $1/\delta$.

Proposition 18. Suppose all weak learners satisfy the weak learning condition with edge γ and sample complexity S (Definition 12). Then with probability at least $1-\delta$, the predictions $(\widehat{y}_t)_{t\le n}$ generated by Algorithm 11 satisfy

$$\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} = \tilde{O}\left(\frac{n}{N\gamma^2} + \frac{1}{\gamma^2} + \frac{KS}{\gamma}\right). \tag{10.7}$$

Thus, to achieve a target error rate ε, it suffices to take $N = \tilde{\Omega}\left(\frac{1}{\varepsilon\gamma^2}\right)$ and $n = \tilde{\Omega}(\frac{1}{\varepsilon\gamma^2} + \frac{KS}{\varepsilon\gamma})$.

Comparison with prior algorithms Compared to (Jung et al., 2017), our sample complexity on n improves the dependence on K (for OnlineMBBM) and also ε and γ (for AdaBoost.OLM), and is in fact optimal according to their lower bound (Theorem 4). Our bound on the number of weak learners, on the other hand, is weaker compared to the non-adaptive algorithm OnlineMBBM (which has a logarithmic dependence on $1/\varepsilon$), but is still much stronger than that of AdaBoost.OLM since it improves the dependence on K from linear to $\log(K)$. Although not stated explicitly, our results also apply to the binary setting considered in (Beygelzimer et al., 2015) and improve the sample complexity of their AdaBoost.OL algorithm to the optimal bound $\tilde{\Omega}(\frac{1}{\varepsilon\gamma^2} + \frac{S}{\varepsilon\gamma})$. Overall, our results significantly reduce the gap between optimal and adaptive online boosting algorithms.

As a final remark, the same technique used here also readily applies to the online boosting setting for the multi-label ranking problem recently studied by Jung and Tewari (2018).

10.7 Detailed Proofs

10.7.1 Proofs from Section 10.2

Lemma 19. The generalized multiclass logisitic loss is $2L$-Lipschitz with respect to ℓ_∞ norm.

Proof. It is straightforward to verify the identity

$$\nabla_z \ell(z, y) = \left(\sum_k y_k\right)\boldsymbol{\sigma}(z) - y.$$

It follows that $\|\nabla_z \ell(z,y)\|_1 \leq \|y\|_1 \|\boldsymbol{\sigma}(z)\|_1 + \|y\|_1 \leq 2L$. By duality, this implies $2L$-Lipschitzness with respect to ℓ_∞. $\square$

Lemma 20. The function $f(x) = \prod_{k\in[d]} x_k^{\alpha_k}$ is concave over $\mathbb{R}^d_+$ whenever $\alpha_k \geq 0\ \forall k$ and $\sum_{k\in[d]} \alpha_k \leq 1$.

Proof. We will prove that the Hessian of f is negative semidefinite. The Hessian can be written as

$$\nabla^2 f(x) = f(x) \cdot G(x),$$

where the matrix $G(x) \in \mathbb{R}^{d\times d}$ is given by $G(x)_{ii} = \alpha_i(\alpha_i - 1)x_i^{-2}$ and $G(x)_{ij} = \alpha_i\alpha_j x_i^{-1} x_j^{-1}$. Since f is nonnegative, it suffices to show that G is negative semidefinite. Using the reparameterization $y_i = x_i^{-1}$ and the notation $\odot$ for the element-wise product, we can write

$$G(y) = (\alpha \odot y)^{\otimes 2} - \text{diag}(\alpha \odot y^2).$$

For any fixed $y \in \mathbb{R}^d_+$ and any $v \in \mathbb{R}^d$, we have

$$\begin{aligned}\langle v, G(y)v\rangle &= \left(\sum_{k=1}^{d} \alpha_k y_k v_k\right)^2 - \sum_{k=1}^{d} \alpha_k y_k^2 v_k^2 \\ &\le \left(\sum_{k=1}^{d} \alpha_k y_k^2 v_k^2\right)\left(\sum_{k=1}^{d} \alpha_k\right) - \sum_{k=1}^{d} \alpha_k y_k^2 v_k^2 \\ &\le 0.\end{aligned}$$

The first inequality above uses Cauchy-Schwarz and the second uses that $\sum \alpha_k \le 1$. $\square$

Proof of Proposition 16. We first show that the generalized multiclass log loss $\ell_{\log}(p, y) := -\sum_{k\in[K]} y_k \log(p_k)$ is $1/L$-mixable over predictions $p \in \Delta_K$ and outcomes $y \in \mathcal{Y}$. Recall that to show η-mixability it is sufficient to demonstrate that ℓ is η-exp-concave with respect to p (e.g. (Cesa-Bianchi and Lugosi, 2006)) for any $y \in \mathcal{Y}$.

Observe that we have

$$e^{-\eta\ell(p,y)} = \prod_{k\in[K]} p_k^{\eta y_k}.$$

When $\eta \le 1/L$, we have $\sum_{k\in[K]} \eta y_k \le 1$. Since $p \in \Delta_K$ and by the definition of $\mathcal{Y}$, Lemma 20 implies the function $p \mapsto \prod_{k\in[K]} p_k^{\eta y_k}$ is concave, which proves the result.

Exp-concavity implies that for any distribution $\tilde{\pi}$ over Δ_K, the predicition $p_{\tilde{\pi}} = \mathbb{E}_{p\sim\tilde{\pi}}[p]$ certifies the inequality

$$\mathop{\mathbb{E}}_{p\sim\tilde{\pi}}[\exp(-\eta\ell_{\log}(p, y))] \le \exp(-\eta\ell_{\log}(p_{\tilde{\pi}}, y)) \quad y \in \mathcal{Y}.$$

Now, turning to the multiclass logistic loss $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ defined as $\ell(z, y) = -\sum_{k\in[K]} y_k \log(\boldsymbol{\sigma}(z)_k)$, let π be any distribution on $\mathbb{R}^K$. Let $\tilde{\pi}$ be the induced distribution on Δ_K via the softmax function, i.e. a sample from $\tilde{\pi}$ is generated by sampling $z \sim \pi$ and computing $p = \boldsymbol{\sigma}(z)$. Then define $z_\pi = \boldsymbol{\sigma}^+(\mathbb{E}_{z\sim\pi}[\boldsymbol{\sigma}(z)])$. Since $\boldsymbol{\sigma}(z_\pi) = \mathbb{E}_{z\sim\pi}[\boldsymbol{\sigma}(z)] = p_{\tilde{\pi}}$

and $\ell(z,y)=\ell_{\log}(\boldsymbol{\sigma}(z),y)$, the above inequality implies that

$$\mathop{\mathbb{E}}_{z\sim\pi}[\exp(-\eta\ell(z,y))]\leq\exp(-\eta\ell(z_\pi,y))\quad y\in\mathcal{Y}.$$

$\square$

Lemma 21. Suppose a strategy $(\tilde{z}_t)_{t\leq n}$ guarantees a regret inequality

$$\sum_{t=1}^{n}\ell(\tilde{z}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\leq\mathbf{R}.$$

Then for $0\leq\mu\leq 1/2$ the strategy $\hat{z}_t:=\boldsymbol{\sigma}^{+}(\mathrm{smooth}_\mu(\boldsymbol{\sigma}(\tilde{z}_t)))$ guarantees

$$\sum_{t=1}^{n}\ell(\hat{z}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\leq\mathbf{R}+2\mu\sum_{t=1}^{n}\|y_t\|_1,$$

and satisfies $\|\hat{z}_t\|_\infty\leq\log(K/\mu)$.

Proof of Lemma 21. We write regret as

$$\begin{aligned}
&\sum_{t=1}^{n}\ell(\hat{z}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)\\
&=\sum_{t=1}^{n}\ell(\tilde{z}_t,y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(x_t),y_t)+\sum_{t=1}^{n}\ell(\hat{z}_t,y_t)-\sum_{t=1}^{n}\ell(\tilde{z}_t,y_t)\\
&\leq\mathbf{R}+\sum_{t=1}^{n}\ell(\hat{z}_t,y_t)-\sum_{t=1}^{n}\ell(\tilde{z}_t,y_t).
\end{aligned}$$

For the last two terms, fix any round t and define $\tilde{p}=\boldsymbol{\sigma}(\tilde{z}_t)$. Since $\boldsymbol{\sigma}(\hat{z}_t)=(1-\mu)\tilde{p}+\mu\mathbb{1}/K$, we have

$$\ell(\hat{z}_t,y_t)-\ell(\tilde{z}_t,y_t)=\sum_{k\in[K]}y_{t,k}\log\left(\frac{\tilde{p}_k}{(1-\mu)\tilde{p}_k+\mu/K}\right)\leq\log\left(\frac{1}{1-\mu}\right)\sum_{k\in[K]}y_{t,k}\leq 2\mu\|y_t\|_1.$$

The last inequality uses that $\log(1/(1-x))\leq 2x$ for $x\leq 1/2$. Summing up over all rounds t gives us the desired regret bound.

To establish boundedness of the predictions, recall that $\boldsymbol{\sigma}_k^{+}(p)=\log(p_k)$. Letting $p=(1-\mu)\,\mathbb{E}_{W\sim P_t}[\boldsymbol{\sigma}(Wx_t)]+\mu\mathbf{1}/K$, it clearly holds that $p_k\geq\mu/K$, and so $|\boldsymbol{\sigma}_k^{+}(p)|\leq\log(K/\mu)$. $\square$

Proof of Theorem 32. Let $\eta = 1/L$. Let $\tilde{z}_t = \boldsymbol{\sigma}^+(\mathbb{E}_{W\sim P_t}[\boldsymbol{\sigma}(Wx_t)])$ — that is, the prediction for the setting $\mu = 0$. We will first establish a regret bound for the case $\mu = 0$, then reduce the general case to it by approximation.

First observe that due to mixability for $\eta \le 1/L$ (from Proposition 16), we have

$$\sum_{t=1}^{n} \ell(\tilde{z}_t, y_t) \le -\frac{1}{\eta}\sum_{t=1}^{n} \log\left(\int_{\mathcal{W}} \exp(-\eta\ell(Wx_t, y_t))dP_t(W)\right).$$

Let $Z_t = \int_{\mathcal{W}} \exp(-\eta\sum_{s=1}^{t} \ell(Wx_s, y_s))dW$ with the convention $Z_0 = \int_{\mathcal{W}} dW$. Using the definition of P_t, the right-hand-side in the displayed equation above is then equal to

$$\begin{aligned}
-\frac{1}{\eta}\sum_{t=1}^{n}\log(Z_t/Z_{t-1}) &= -\frac{1}{\eta}\log(Z_n/Z_0) \\
&= -\frac{1}{\eta}\log\left(\int_{\mathcal{W}}\exp\left(-\eta\sum_{t=1}^{n}\ell(Wx_t, y_t)\right)dW\right) + \frac{1}{\eta}\log(\mathrm{Vol}(\mathcal{W}))
\end{aligned}$$

We will upper bound the term $-\log\big(\int_{\mathcal{W}}\exp(-\eta\sum_{t=1}^{n}\ell(Wx_t, y_t))dW\big)$. Let $W^\star = \arg\min_{W\in\mathcal{W}}\sum_{t=1}^{n}\ell(Wx_t, y_t)$. Fix $\theta \in [0, 1)$ and let $S = \{\theta W^\star + (1-\theta)W \mid W \in \mathcal{W}\} \subseteq \mathcal{W}$. To upper bound the negative-log-integral term, we will lower bound the integral appearing inside.

$$\int_{\mathcal{W}}\exp\left(-\eta\sum_{t=1}^{n}\ell(Wx_t, y_t)\right)dW \ge \int_{S}\exp\left(-\eta\sum_{t=1}^{n}\ell(Wx_t, y_t)\right)dW.$$

Using a change of variables and noting that since $W \in \mathbb{R}^{K\times d}$ the Jacobian of the mapping $W \mapsto (1-\theta)W + \theta W^\star$ has determinant $(1-\theta)^{D_{\mathcal{W}}}$, the right-hand-side above equals

$$= (1-\theta)^{D_{\mathcal{W}}}\int_{\mathcal{W}}\exp\left(-\eta\sum_{t=1}^{n}\ell((\theta W^\star + (1-\theta)W)x_t, y_t)\right)dW.$$

Observe that $\|(\theta W^\star + (1-\theta)W)x_t - W^\star x_t\|_\infty = (1-\theta)\max_{k\in[K]}|\langle W_k^\star - W_k, x_t\rangle| \leq 2(1-\theta)B\|x_t\|_\star$. Using this observation with the $2L$-Lipschitzness of ℓ with respect to ℓ_∞ from Lemma 19 implies that the above displayed expression is at most

$$
\begin{aligned}
&(1-\theta)^{D_\mathcal{W}} \int_\mathcal{W} \exp\left(-\eta \sum_{t=1}^n \ell(W^\star x_t, y_t) - 4(1-\theta)BL\eta \sum_{t=1}^n \|x_t\|_\star\right) dW. \\
&= (1-\theta)^{D_\mathcal{W}} \cdot \mathrm{Vol}(\mathcal{W}) \cdot \exp\left(-\eta \sum_{t=1}^n \ell(W^\star x_t, y_t)\right) \cdot \exp\left(-4(1-\theta)BL\eta \sum_{t=1}^n \|x_t\|_\star\right).
\end{aligned}
$$

Combining all of the observations so far, we have proven the following regret bound:

$$
\begin{aligned}
&\sum_{t=1}^n \ell(\widehat{y}_t, y_t) - \sum_{t=1}^n \ell(W^\star x_t, y_t) \\
&\leq \frac{1}{\eta}\log(\mathrm{Vol}(\mathcal{W})) - \sum_{t=1}^n \ell(W^\star x_t, y_t) \\
&\quad + \frac{1}{\eta}\underbrace{\left(D_\mathcal{W}\log\left(\frac{1}{1-\theta}\right) - \log(\mathrm{Vol}(\mathcal{W})) + \eta\sum_{t=1}^n \ell(W^\star x_t, y_t) + 4(1-\theta)BL\eta\sum_{t=1}^n \|x_t\|_\star\right)}_{\text{Bound on negative log-integral-exp.}} \\
&= \frac{D_\mathcal{W}}{\eta}\log\left(\frac{1}{1-\theta}\right) + 4(1-\theta)BL\sum_{t=1}^n \|x_t\|_\star.
\end{aligned}
$$

To conclude, we choose θ to satisfy $1-\theta = \min\{D_\mathcal{W}/(B\sum_{t=1}^n \|x_t\|_\star), 1\}$. Note that regardless of which argument obtains the minimum, we have $4(1-\theta)BL\sum_{t=1}^n \|x_t\|_\star \leq 4D_\mathcal{W}L$. The choice of θ also means that $\log(\frac{1}{1-\theta}) = \log(1 \vee B\sum_{t=1}^n \|x_t\|_\star / D_\mathcal{W})$. This leads to a final bound of

$$
D_\mathcal{W}L \cdot \log\left(1 \vee \frac{B\sum_{t=1}^n \|x_t\|_\star}{D_\mathcal{W}}\right) + 4D_\mathcal{W}L.
$$

To simplify we upper bound this by

$$
5D_\mathcal{W}L \cdot \log\left(\frac{B\sum_{t=1}^n \|x_t\|_\star}{D_\mathcal{W}} + e\right) = 5D_\mathcal{W}L \cdot \log\left(\frac{BRn}{D_\mathcal{W}} + e\right).
$$

To handle the general case where $\mu > 0$ we simply appeal to Lemma 21 and use that $\boldsymbol{\sigma}(\boldsymbol{\sigma}^+(p)) = p \; \forall p \in \Delta_K$.

$\square$

We now state the proof of Theorem 33. This proof is a simple corollary of Theorem 38, a lower bound on mistakes for online binary classification with a margin. Theorem 38 is proven in the remainder of this section of the appendix. To begin, we need the following definition:

Definition 13. *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be some function class. A dataset $(x_1, y_1), \dots, (x_n, y_n) \in \cup_{t=1}^{n} \mathcal{X} \times \{\pm 1\}$ is shattered with γ margin if there exists $f \in \mathcal{F}$ such that*

$$f(x_t)y_t \geq \gamma.$$

Proof of Theorem 33. Let $\hat{z}_t$ for $t \in [n]$ be the sequence of predictions made by the algorithm for a sequence of examples (x_t, y_t), for $t \in [n]$. It is easy to check that

$$\sum_{t=1}^{n} \ell_{\text{bin}}(\hat{z}_t, y_t) \geq \log(2) \sum_{t=1}^{n} \mathbb{1}\{\text{sgn}(\hat{z}_t) \neq y_t\}.$$

Let $1/\gamma = B/\log(n)$. From Theorem 38, it holds that whenever $\gamma \leq O(1/\sqrt{d})$, there exists an adversarial sequence (x_t, y_t), for $t \in [n]$, for which

$$\sum_{t=1}^{n} \mathbb{1}\{\text{sgn}(\widehat{y}_t) \neq y_t\} \geq \frac{d}{4} \left\lfloor \log_2\left(\frac{1}{5\gamma d^{1/2}}\right) \right\rfloor,$$

and for which the dataset is γ-shattered by some $w \in \mathbb{R}^d$ with $\|w\|_2 \leq 1$. Since the dataset is γ-shattered we also have

$$\inf_{w:\|w\|_2 \leq B} \sum_{t=1}^{n} \ell_{\text{bin}}(\langle w, x_t \rangle, y_t) \leq \sum_{t=1}^{n} \log(1 + e^{-\gamma B}) = \sum_{t=1}^{n} \log\left(1 + \frac{1}{n}\right) \leq 1.$$

This yields the desired lower bound on the regret. □

Theorem 38. *Fix a margin $\gamma \in (0, \frac{1}{4\sqrt{5d}}]$. Then for any randomized strategy $(\widehat{y}_t)_{t \leq n}$ there exists an adversary $(x_t)_{t \leq n}$, $(y_t)_{t \leq n}$ with $\|x_t\|_2 \leq 2$ for which*

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{\text{sgn}(\widehat{y}_t) \neq y_t\}\right] \geq \frac{d}{4} \left\lfloor \log_2\left(\frac{1}{5\gamma d^{1/2}}\right) \right\rfloor, \tag{10.8}$$

and the data sequence is realizable by a unit vector $w \in \mathbb{R}^{d+1}$ with margin γ.

Remark 3. *This lower bound only applies in the regime where $\frac{1}{\gamma^2} \geq d$, meaning that it does not contradict the dimension-independent Perceptron bound.*

To prove Theorem 38, we first state a standard lower bound based on Littlestone's dimension.

Definition 14. *An $\mathcal{X}$-valued tree is a sequence of mappings $\boldsymbol{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$ for $1 \leq t \leq n$.*

We use the abbreviation of $\boldsymbol{x}_t(\epsilon) = \boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1})$ for such a tree, where $\epsilon \in \{\pm 1\}^n$.

Lemma 22. Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be some function class. Suppose there exists a $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth D_γ such that

$$\forall \epsilon \in \{\pm 1\}^{D_\gamma} \;\; \exists f \in \mathcal{F} \quad \text{s.t.} \quad f(\boldsymbol{x}_t(\epsilon))\epsilon_t \geq \gamma. \tag{10.9}$$

Then

$$\inf_{q_1, \ldots, q_n} \sup_{\substack{(x_1, y_1), \ldots, (x_n, y_n) \\ \text{separable with } \gamma \text{ margin}}} \mathop{\mathbb{E}}_{\widehat{y}_1 \sim q_t, \ldots, \widehat{y}_n \sim q_n} \left[\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} \right] \geq \frac{1}{2} \min\{D_\gamma, n\},$$

where the infimum and supremum above are understood to range over policies.

Proof of Lemma 22. Suppose that $n \leq D_\gamma$. We will sample Rademacher random variables $\epsilon \in \{\pm 1\}^n$ and play $y_t = \epsilon_t$ and $x_t = \boldsymbol{x}_t(\epsilon_{1:t-1})$. This immediately implies that the expected number of mistakes is equal to $\frac{n}{2}$. Moreover, since $n \leq D_\gamma$, the assumption in the statement of the lemma implies that there exists $f \in \mathcal{F}$ such that $f(\boldsymbol{x}_t(\epsilon))y_t \geq \gamma$, so the data is indeed separable with γ margin.

If $n > D_\gamma$ we can follow the strategy above, then continue to play $(x_{D_\gamma}, y_{D_\gamma})$ for all $t > D_\gamma$. □

Proof of Theorem 38. By Lemma 22 it suffices to exhibit a tree $\boldsymbol{x}$ for which (10.9) is satisfied with $D_\gamma = \Omega(d \log(1/(\sqrt{d}\gamma)))$.

We first restate a well-known tree instance for the one-dimensional case. Consider a class of thresholds $\mathcal{F}_{\text{thresh}} = \{f_\theta : [0,1] \to \{\pm 1\}\}$ defined by $f_\theta(z) = 1 - 2\mathbb{1}\{x < \theta\}$. The claim is as follows: For any $\delta \in (0,1]$, there exists a $[0,1]$-valued tree $\boldsymbol{z}$ of depth $D_\delta := \lfloor \log_2(2/\delta) \rfloor$ such that

1. $\forall \epsilon \in \{\pm 1\}^{D_\delta} \;\; \exists \theta \quad \text{s.t.} \quad f_\theta(\boldsymbol{z}_t(\epsilon))\epsilon_t = 1.$
2. $|\boldsymbol{z}_t(\epsilon) - \boldsymbol{z}_s(\epsilon)| \geq \delta \;\; \forall s \neq t.$

The construction is as follows. Let $u_1 = 1$, $l_1 = 0$. Recursively for $t = 1, \ldots, n$:

- $\boldsymbol{z}_t(\epsilon_{1:t-1}) = \frac{l_t + u_t}{2}$.
- If $\epsilon_t = -1$ set $l_{t+1} = \boldsymbol{z}_t(\epsilon_{1:t-1})$ and $u_{t+1} = u_t$, else set $u_{t+1} = \boldsymbol{z}_t(\epsilon_{1:t-1})$ and $l_{t+1} = l_t$.

Under this construction the sequence $\boldsymbol{z}_1(\cdot), \ldots, \boldsymbol{z}_{D_\delta}(\epsilon_{1:D_\delta - 1})$ can always be shattered. Furthermore $\boldsymbol{z}^\star(\epsilon) := \boldsymbol{z}_{D_\delta + 1}(\epsilon_{1:D_\delta})$ satisfies the additional property that $\boldsymbol{z}_t > \boldsymbol{z}^\star(\epsilon) \implies \epsilon_t = 1$ and $\boldsymbol{z}_t < \boldsymbol{z}^\star(\epsilon) \implies \epsilon_t = -1$. Also, $|\boldsymbol{z}^\star - \boldsymbol{z}_t| \geq \frac{\delta}{2} \; \forall t \leq D_\delta$.

We now show how to extend this instance to $d+1$ dimensions for any $d \geq 1$. The approach is to concatenate d instances of the $\boldsymbol{z}$ tree constructed above, one for each of the first d coordinates. The final coordinate is left as a constant so that a bias can be implemented.

Let $n = d \cdot D_\delta$ be the tree depth for our $d+1$-dimensional instance. For any time t, let $k \in [d]$ and $\tau \in [D_\delta]$ be such that $t = (k-1)D_\delta + \tau$. Let any sequence $\epsilon \in \{\pm 1\}^n$ be partitioned as $(\boldsymbol{\epsilon}^1, \ldots, \boldsymbol{\epsilon}^d)$ with each $\boldsymbol{\epsilon}^k \in \{\pm 1\}^{D_\delta}$. Letting e_k denote the kth standard basis vector, we define a shattered tree $\boldsymbol{x}$ as follows:

$$\boldsymbol{x}_t(\epsilon_{1:t-1}) = e_{d+1} + e_k \boldsymbol{z}_\tau(\boldsymbol{\epsilon}^k_{1:\tau-1}).$$

We construct a vector $w \in \mathbb{R}^{d+1}$ whose sign correctly classifies each $\boldsymbol{x}_t$ as follows:

- $w_{d+1} = -\delta$.
- $w_k = \delta / \boldsymbol{z}^\star(\boldsymbol{\epsilon}^k)$.

For any $t = (k-1)D_\delta + \tau$ this choice gives

$$\langle w, \boldsymbol{x}_t(\epsilon)\rangle \epsilon_t = \delta(\boldsymbol{z}_\tau(\boldsymbol{\epsilon}^k_{1:\tau-1})/\boldsymbol{z}^\star(\boldsymbol{\epsilon}^k) - 1)\epsilon_t.$$

As described above, $\boldsymbol{z}_t > \boldsymbol{z}^\star(\epsilon) \implies \epsilon_t = 1$ and $\boldsymbol{z}_t < \boldsymbol{z}^\star(\epsilon) \implies \epsilon_t = -1$, which immediately implies that the inner product is always non-negative, and so the dataset is shattered. Using that $|\boldsymbol{z}^\star(\epsilon) - \boldsymbol{z}_t(\epsilon)| \geq \frac{\delta}{2}$ and that both numbers lie in $[0,1]$, we can lower bound the magnitude with which the shattering takes place:

$$\left|\boldsymbol{z}_\tau(\epsilon^k_{1:\tau-1})/\boldsymbol{z}^\star(\boldsymbol{\epsilon}^k) - 1\right| = \frac{1}{\boldsymbol{z}^\star(\boldsymbol{\epsilon}^k)}\left|\boldsymbol{z}_\tau(\epsilon^k_{1:\tau-1}) - \boldsymbol{z}^\star(\boldsymbol{\epsilon}^k)\right| \geq \frac{1}{\boldsymbol{z}^\star(\boldsymbol{\epsilon}^k)}\frac{\delta}{2} \geq \frac{\delta}{4},$$

and so the shattering takes place with margin at least $\delta^2/4$.

Lastly, the norm of w is given by

$$\|w\|_2 = \sqrt{\delta^2 + \sum_{k=1}^{d}\left(\frac{\delta}{\boldsymbol{z}^\star(\boldsymbol{\epsilon}^k)}\right)^2} \leq \sqrt{\delta^2 + 4d} \leq \sqrt{5d},$$

where the first inequality uses that $\boldsymbol{z}^\star(\epsilon) \geq \delta/2$ and the second uses that $d \geq 1$

Rescaling, we have that the vector $w/\|w\|_2$ shatters the tree with margin at least $\frac{\delta^2}{4\sqrt{5d}}$. To rephrase the result as a function of a desired margin: For any margin $\gamma \in (0, \frac{1}{4\sqrt{5d}}]$, setting $\delta = \sqrt{\gamma 4\sqrt{5d}} \leq 1$, we have constructed a tree of depth $\left\lfloor \log_2(2/\sqrt{\gamma 4\sqrt{5d}})\right\rfloor$ that can be shattered with margin γ.

$\square$

10.7.2 Proof from Section 10.3

Theorem 39. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \Delta_K$. Suppose there is an online multiclass learning algorithm over $\mathcal{F}$ using the log loss that for any data sequence $(x_t, y_t) \in \mathcal{X} \times [K]$ for $t = 1, 2, \ldots, n$ produces distributions $p_t \in \Delta_K$ such that the following regret bound holds:*

$$\sum_{t=1}^{n} \ell_{\log}(p_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell_{\log}(f(x_t), y_t) \leq R(n).$$

Here $R(n)$ is some function of n and other relevant problem dependent parameters. Then for any given $\delta > 0$ and any (unknown) distribution $\mathcal{D}$ over $\mathcal{X} \times [K]$, it is possible to construct a predictor $g : \mathcal{X} \to \Delta_K$ using n samples $\{(x_t, y_t)\}_{t=1}^{n}$ drawn from $\mathcal{D}$ such that with probability at least $1 - \delta$, the excess risk of g is bounded as

$$\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(g(x), y)] \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(f(x), y)] + O\left(\frac{\log\left(\frac{1}{\delta}\right) R\left(\frac{n}{\log(1/\delta)}\right) + \log(Kn) \log\left(\frac{\log(n)}{\delta}\right)}{n}\right).$$

Proof of Theorem 39. Recall that the standard online-to-batch conversion (Helmbold and Warmuth, 1995) produces an (improper) predictor using n data samples by running the online algorithm on those samples and stopping at a random time. Then predictor is online algorithm with its the internal state frozen. This predictor has excess risk bounded by the average regret over n rounds, in expectation over the n data samples.

The algorithm to generate the predictor g with the specified excess risk bound in the theorem statement is given below:

1. Let $M = \lceil \log(2/\delta) \rceil$. Produce M predictors $h_1, \ldots, h_M : \mathcal{X} \to \Delta_K$ by using the online-to-batch conversion on the online multiclass learning algorithm run using M disjoint sets of $n/2M$ samples each. Call the ith such set of samples S_i
2. For $i \in [M]$, define $\tilde{h}_i : \mathcal{X} \to \Delta_K$ as $\tilde{h}_i(x) = \mathrm{smooth}_\mu(h_i(x))$ for $\mu = \frac{R(n/M)}{2n/M}$.

3. Construct an online convex optimization instance as follows. The learner's decision set is Δ_M, the set of all distributions on $[M]$. For every data point $(x, y) \in \mathcal{X} \times [K]$, associate the loss function $\ell_{(x,y)} : \Delta_M \to \mathbb{R}$ defined as $\ell_{(x,y)}(q) = -\log(\mathbb{E}_{i\sim q}[(\tilde{h}_i(x))_y])$. These loss functions are 1-exp-concave, so run the EWOO algorithm (Hazan et al., 2007) using the remaining $n/2$ examples sequentially to generate loss functions. Let $\bar{q}$ be the average of all the distributions in Δ_M generated by EWOO. Define $g := \mathbb{E}_{i\sim\bar{q}}[\tilde{h}_i]$.

We now proceed to analyse the excess risk of g. First, using the regret bound for the online multiclass learning algorithm, and in-expecation bound on the excess risk for online-to-batch conversion, for every $i \in [M]$, we have

$$\mathop{\mathbb{E}}_{S_i}\left[\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(h_i(x), y)]\right] \leq \inf_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(f(x), y)] + \frac{R(n/M)}{n/M}.$$

For any $p \in \Delta_K$, if $\tilde{p} = \mathrm{smooth}_\mu(p)$, then for any $y \in [K]$ we have $-\log(\tilde{p}_y) + \log(p_y) = \log(\frac{p_y}{(1-\mu)p_y+\mu/K}) \leq 2\mu$. So for every $i \in [M]$, we have

$$\mathop{\mathbb{E}}_{S_i}\left[\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(\tilde{h}_i(x), y)]\right] \leq \mathop{\mathbb{E}}_{S_i}\left[\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(h_i(x), y)]\right] + 2\mu.$$

Putting the above two bounds together, using the specified value of μ and an application of Markov's inequality, with probability at least $1 - e^{-M} = 1 - \frac{\delta}{2}$, there exists some $i^\star \in [M]$ such that

$$\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(\tilde{h}_{i^\star}(x), y)] \leq \inf_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(f(x), y)] + \frac{2eR(n/M)}{n/M}. \tag{10.10}$$

The EWOO algorithm in step 3 of the procedure enjoys a regret bound of $O(M\log(n))$ (the online convex optimization problem is an instance of online portfolio selection over M instruments, see (Hazan et al., 2007)). Furthermore, the application of smooth_μ makes the range for the log loss be bounded by $\log(K/\mu)$. Thus, by Corollary 2 of Mehta (2017), with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(g(x),y)] &= \mathop{\mathbb{E}}_{(x,y)}[-\log(\mathop{\mathbb{E}}_{i\sim\bar{q}}[(\tilde{h}_i(x))_y])] \\ &\leq \mathop{\mathbb{E}}_{(x,y)}[-\log((\tilde{h}_{i^\star}(x))_y)] + O\left(\frac{M\log(n) + \log(K/\mu)\log(\log(n)/\delta)}{n}\right)\end{aligned} \tag{10.11}$$

Note that $\ell_{\log}(\tilde{h}_{i^\star}(x),y) = -\log((\tilde{h}_{i^\star}(x))_y)$. Applying the union bound and combining inequalities (10.10) and (10.11) with some simplification of the bounds using the value of M, with probability at least $1-\delta$ we have

$$\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(g(x),y)] \leq \inf_{f\in\mathcal{F}}\mathop{\mathbb{E}}_{(x,y)}[\ell_{\log}(f(x),y)] + O\left(\frac{\log\left(\frac{1}{\delta}\right)R\left(\frac{n}{\log(1/\delta)}\right) + \log(Kn)\log\left(\frac{\log(n)}{\delta}\right)}{n}\right).$$

□

10.7.3 Proofs from Section 10.7.3

For this section we let ℓ denote the unweighted multiclass logistic loss: the multiclass logistic loss defined in Section 10.1.1 for the special case where $\mathcal{Y} = \{e_i\}_{i\in[K]}$. Before proving Theorem 35 we need a few preliminaries. First, we state a version of the Aggregating Algorithm with the logistic loss for finite classes.

Lemma 23. Let $\mathcal{F}$ be any finite class of sequences of the form $f = (f_t)_{t\leq n}$ with $f_t \in \mathbb{R}^K$, where each f_t is available at time t and may depend on $y_{1:t-1}$. Define a strategy

1. $P_t(f) \propto \exp\left(-\sum_{s=1}^{t-1}\ell(f_s,y_s)\right)$ (so $P_1 = \mathrm{Uniform}(\mathcal{F})$).
2. $\hat{z}_t = \boldsymbol{\sigma}^+(\mathrm{smooth}_{\frac{1}{n}}(\mathbb{E}_{f\sim P_t}[\boldsymbol{\sigma}(f_t)]))$.

This strategy enjoys a regret bound of

$$\sum_{t=1}^{n}\ell(\hat{z}_t,y_t) - \min_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f_t,y_t) \leq \log|\mathcal{F}| + 2. \tag{10.12}$$

Furthermore, the predictions satisfy $\|\hat{z}_t\|_\infty \leq \log(Kn)$.

Proof of Lemma 23. First consider the closely related strategy $\widetilde{z}_t := \boldsymbol{\sigma}^+(\mathbb{E}_{f\sim P_t}[\boldsymbol{\sigma}(f(x_t))])$. In light of the 1-mixability for the logistic loss proven in Proposition 15, $\widetilde{z}_t$ is precisely the finite class version of the Aggregating Algorithm, which guarantees (Cesa-Bianchi and Lugosi, 2006):

$$\sum_{t=1}^{n} \ell(\widetilde{z}_t, y_t) - \min_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f_t, y_t) \leq \log|\mathcal{F}|.$$

To establish the final result we simply appeal to Lemma 21, using that $\boldsymbol{\sigma}(\boldsymbol{\sigma}^+(p)) = p \; \forall p \in \Delta_K$. $\square$

We require need a slight generalization of the notion of covering number defined in Definition 17 for intermediate results.

Definition 15. *Let U be a collection of $\mathbb{R}^K$-valued K-ary trees. A set V of $\mathbb{R}^K$-valued K-ary trees is an α-cover with respect to the L_p norm for U if*

$$\forall \mathbf{u} \in U,\; y \in [K]^n,\; \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{n} \max_{y_t'\in[K]} |\ell(\mathbf{u}_t(y), y_t') - \ell(\mathbf{v}_t(y), y_t')|^p\right)^{1/p} \leq \alpha.$$

Definition 16. *The L_p covering number for a collection of trees U with loss ℓ is*

$$\mathcal{N}_p(\alpha, \ell\circ U) := \min\{|V| : V \textit{ is an } \alpha\text{-cover of } U \text{ w.r.t. the } L_p \text{ norm}\}.$$

Proof of Theorem 35. Define a subset of the output space:

$$\mathcal{Z} := \left\{z \in \mathbb{R}^K \mid \|z\|_\infty \leq \log(Kn)\right\}.$$

We move to an upper bound on the minimax value by restricting predictions to $\mathcal{Z}$:

$$\begin{aligned}
\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}) &= \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \inf_{\hat{z}_t\in\mathbb{R}^K} \max_{y_t\in[K]} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right] \\
&\leq \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \inf_{\hat{z}_t\in\mathcal{Z}} \max_{y_t\in[K]} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right].
\end{aligned}$$

Note that $\mathcal{Z}$ is a compact subset of a separable metric space and that ℓ is convex with respect to $\hat{z}$. Therefore, using repeated application of minimax theorem following Section 2.6, the minimax value can be written as:

$$= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_K} \inf_{\hat{z}_t \in \mathcal{Z}} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right].$$

Now we perform a standard manipulation of the sup and loss terms as in Rakhlin et al. (2010):

$$= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{p_t \in \Delta_K} \mathbb{E}_{y_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \inf_{\hat{z}_t \in \mathcal{Z}} \mathbb{E}_{y_t \sim p_t} \left[\ell(\hat{z}_t, y_t)\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \tag{10.13}$$

$$= \sup_{\mathbf{x}, \mathbf{p}} \mathbb{E}_{y \sim \mathbf{p}} \left[\sum_{t=1}^{n} \inf_{\hat{z}_t \in \mathcal{Z}} \mathbb{E}_{y_t \sim \mathbf{p}_t(y)} \left[\ell(\hat{z}_t, y_t)\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(\mathbf{x}_t(y)), y_t) \right]. \tag{10.14}$$

In the final line above we have introduced new notation. $\mathbf{x}$ and $\mathbf{p}$ are $\mathcal{X}$- and Δ_K-valued K-ary trees of depth n. That is, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_t : [K]^{t-1} \to \mathcal{X}$ and similarly for the tree $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$, $\mathbf{p}_t : [K]^{t-1} \to \Delta_K$. The notation "$y \sim \mathbf{p}$" refers to the process in which we first draw $y_1 \sim \mathbf{p}_1$, then draw $y_t \sim \mathbf{p}_t(y_1, \ldots, y_{t-1})$ for subsequent timesteps t. We also overload the notation as $\mathbf{p}_t(y) := \mathbf{p}_t(y_{1:t-1})$, and likewise for $\mathbf{x}$.

With this notation, (10.14) is seen to be (10.13) rewritten using that at time t, based on draw of previous ys, x_t and p_t are chosen to maximize the remaining game value; this process be represented via K-ary tree.

Note that the sequence $(\hat{z}_t)_{t \le n}$ being minimized over in (10.13) can depend on the full trees $\mathbf{x}$ and $\mathbf{p}$, but that it is adapted to the path $(y_t)_{t \le n}$, meaning that the value at time t ($\hat{z}_t$) can only depend on the $y_{1:t-1}$. This property is imporant because the choice we exhibit for $(\hat{z}_t)_{t \le n}$ will indeed depend on the full trees.

In light of the discussion in Section 10.7.3, the key advantage of having moved to the dual game above is that we can condition on the K-ary tree $\mathbf{x}$ and cover $\mathcal{F}$ only on this tree. Let

V^γ be a minimal γ-sequential cover of $\ell \circ \mathcal{F}$ on the tree $\mathbf{x}$ with respect to the L_2 norm (in the sense of Definition 17).

Keeping the tree $\mathbf{x}$ fixed, for each tree $\mathbf{v} \in V^\gamma$, each $f \in \mathcal{F}$, we define a class of trees $\mathcal{F}_{\mathbf{v}}$ "centered" at $\mathbf{v}$—in a sense that will be made precise in a moment—via the following procedure.

- $\mathcal{F}_{\mathbf{v}} = \emptyset$.
- For each $f \in \mathcal{F}$ and $y \in [K]^n$ with $\sqrt{\frac{1}{n}\sum_{t=1}^{n} \max_{y_t'' \in [K]}(\ell(f(\mathbf{x}_t(y)), y_t'') - \ell(\mathbf{v}_t(y), y_t''))^2} \leq \gamma$:
  - Define a $\mathbb{R}^K$-valued K-ary tree $\mathbf{u}_{f,y}$ via: For each $y' \in [K]^n$,

$$(\mathbf{u}_{f,y})_t(y') \\ := f(\mathbf{x}_t(y'))\mathbb{1}\{y_1' = y_1, \ldots, y_{t-1}' = y_{t-1}\} + \mathbf{v}_t(y')\mathbb{1}\{\neg(y_1' = y_1, \ldots, y_{t-1}' = y_{t-1})\}.$$

    In other words, $\mathbf{u}_{f,y}$ is equal to $f \circ \mathbf{x}$ on the path y, and equal to $\mathbf{v}$ everywhere else.
  - Add $\mathbf{u}_{f,y}$ to $\mathcal{F}_{\mathbf{v}}$.

The class $\mathcal{F}_{\mathbf{v}}$ has two important properties which are formally proven in an auxiliary lemma, Lemma 24: First, its L_2 covering number is (up to low order terms) bounded in terms of the L_2 covering number of the class $\mathcal{F} \circ \mathbf{x}$, so it has similar complexity to this class. Second, its L_2 radius is bounded by γ, in the sense that its covering number at scale γ is at most 1.

Note that on any path $y \in [K]^n$ and for each $f \in \mathcal{F}$, there exist $\mathbf{v} \in V^\gamma$ and $\mathbf{u} \in \mathcal{F}_{\mathbf{v}}$ such that $f(\mathbf{x}_t(y)) = \mathbf{u}_t(y)$. This is because a $\mathbf{v}$ that is γ-close to f on the path y through $\mathbf{x}$ is guaranteed by the cover property of V^γ, and so we can take $\mathbf{u}_{f,y}$ in $\mathcal{F}_{\mathbf{v}}$ as the desired $\mathbf{u}$.

This implies that

$$\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(\mathbf{x}_t(y)),y_t)\geq\min_{\mathbf{v}\in V^\gamma}\inf_{\mathbf{u}\in\mathcal{F}_\mathbf{v}}\sum_{t=1}^{n}\ell(\mathbf{u}_t(y),y_t).$$

With this we are ready to return to the minimax rate. We already established that

$$\mathcal{V}_n(\mathcal{F})\leq\sup_{\mathbf{x},\mathbf{p}}\mathbb{E}_{y\sim\mathbf{p}}\left[\sum_{t=1}^{n}\inf_{\hat{z}_t\in\mathcal{Z}}\mathbb{E}_{y_t\sim\mathbf{p}_t(y)}\left[\ell(\hat{z}_t,y_t)\right]-\inf_{f\in\mathcal{F}}\sum_{t=1}^{n}\ell(f(\mathbf{x}_t(y)),y_t)\right].$$

We now move to an upper bound based on the constructions for the tree collections V^γ and $\{\mathcal{F}_\mathbf{v}\}_{\mathbf{v}\in V^\gamma}$. These collections depend only on the tree $\mathbf{x}$ at the outer supremum above. Writing the choice of these collections as an infimum to make its dependence on the other quantities in the random process as explicit as possible, and using the containment just shown:

$$\leq\sup_{\mathbf{x}}\inf_{V^\gamma}\inf_{\{\mathcal{F}_\mathbf{v}\}_{\mathbf{v}\in V^\gamma}}\sup_{\mathbf{p}}\mathbb{E}_{y\sim\mathbf{p}}\left[\sum_{t=1}^{n}\inf_{\hat{z}_t\in\mathcal{Z}}\mathbb{E}_{y_t\sim\mathbf{p}_t(y)}\left[\ell(\hat{z}_t,y_t)\right]-\min_{\mathbf{v}\in V^\gamma}\inf_{\mathbf{u}\in\mathcal{F}_\mathbf{v}}\sum_{t=1}^{n}\ell(\mathbf{u}_t(y),y_t)\right].$$

For the last time in the proof, we introduce a new collection of trees. For each $\mathbf{v}\in V^\gamma$ we introduce a $\mathcal{Z}$-valued K-ary tree $\hat{\mathbf{y}}^\mathbf{v}$, with $\hat{\mathbf{y}}^\mathbf{v}_t:[K]^{t-1}\to\mathcal{Z}$. We postpone explicitly constructing the trees for now, but the reader may think of each tree $\hat{\mathbf{y}}^\mathbf{v}$ as representing the optimal strategy for the set $\mathcal{F}_\mathbf{v}$ in a sense that will be made precise in a moment.

$$\begin{aligned}&=\sup_{\mathbf{x}}\inf_{V^\gamma}\inf_{\{\mathcal{F}_\mathbf{v}\}_{\mathbf{v}\in V^\gamma}}\inf_{\{\hat{\mathbf{y}}^\mathbf{v}\}_{\mathbf{v}\in V^\gamma}}\sup_{\mathbf{p}}\mathbb{E}_{y\sim\mathbf{p}}\Bigg[\sum_{t=1}^{n}\inf_{\hat{z}_t\in\mathcal{Z}}\mathbb{E}_{y_t\sim\mathbf{p}_t(y)}\left[\ell(\hat{z}_t,y_t)\right]\\&\qquad-\min_{\mathbf{v}\in V^\gamma}\Bigg\{\sum_{t=1}^{n}\ell(\hat{\mathbf{y}}^\mathbf{v}_t(y),y_t)\\&\qquad\qquad-\sum_{t=1}^{n}\ell(\hat{\mathbf{y}}^\mathbf{v}_t(y),y_t)+\inf_{\mathbf{u}\in\mathcal{F}_\mathbf{v}}\sum_{t=1}^{n}\ell(\mathbf{u}_t(y),y_t)\Bigg\}\Bigg].\\&\leq\sup_{\mathbf{x}}\inf_{V^\gamma}\inf_{\{\mathcal{F}_\mathbf{v}\}_{\mathbf{v}\in V^\gamma}}\inf_{\{\hat{\mathbf{y}}^\mathbf{v}\}_{\mathbf{v}\in V^\gamma}}\left\{\begin{array}{c}\underbrace{\sup_{\mathbf{p}}\mathbb{E}_{y\sim\mathbf{p}}\left[\sum_{t=1}^{n}\inf_{\hat{z}_t\in\mathcal{Z}}\mathbb{E}_{y_t\sim\mathbf{p}_t(y)}\left[\ell(\hat{z}_t,y_t)\right]-\min_{\mathbf{v}\in V^\gamma}\sum_{t=1}^{n}\ell(\hat{\mathbf{y}}^\mathbf{v}_t(y),y_t)\right]}_{(\star)}\\+\underbrace{\sup_{\mathbf{p}}\mathbb{E}_{y\sim\mathbf{p}}\left[\max_{\mathbf{v}\in V^\gamma}\left\{\sum_{t=1}^{n}\ell(\hat{\mathbf{y}}^\mathbf{v}_t(y),y_t)-\inf_{\mathbf{u}\in\mathcal{F}_\mathbf{v}}\sum_{t=1}^{n}\ell(\mathbf{u}_t(y),y_t)\right\}\right]}_{(\star\star)}\end{array}\right\}.\end{aligned}\tag{10.15}$$

We now bound the terms $(\star)$ and $(\star\star)$ individually by instantiating specific choices for $(\hat{z}_t)_{t\leq n}$ and $\{\hat{\mathbf{y}}^{\mathbf{v}}\}$.

Term $(\star)$ We select $(\hat{z}_t)_{t\leq n}$ using the Aggregating Algorithm as configured in Lemma 23, taking $\mathcal{F}$ to be the finite collection of sequences $\{\hat{\mathbf{y}}^{\mathbf{v}}\}_{\mathbf{v}\in V^\gamma}$. Since each tree has the property that $\hat{\mathbf{y}}_t^{\mathbf{v}}$ only depends on $y_{1:t-1}$, Lemma 23 indeed applies, which means that for any sequence $y_{1:n}\in[K]^n$ of labels the algorithm deterministically satisfies the regret inequality

$$\sum_{t=1}^{n}\ell(\hat{z}_t,y_t)-\min_{\mathbf{v}\in V^\gamma}\sum_{t=1}^{n}\ell(\hat{\mathbf{y}}_t^{\mathbf{v}}(y),y_t)\leq\ \log|V^\gamma|+2.$$

Since the algorithm guarantees $\|\hat{z}_t\|_\infty\leq\ \log(Kn)$, one can verify that $\hat{z}_t\in\mathcal{Z}$. Furthermore, $\hat{z}_t$ depends only on $y_{1:t-1}$, and so the predictions of the Aggregating Algorithm are a valid choice for the infimum in $(\star)$. This implies that

$$(\star)\leq\ \sup_{\mathbf{x}}\log|V^\gamma|+2\leq\log\mathcal{N}_2(\gamma,\ell\circ\mathcal{F})+2,$$

since the regret inequality holds for every possible draw of $y_{1:n}$ in the expression $(\star)$.

Term $(\star\star)$ First, observe that each tree class $\mathcal{F}_{\mathbf{v}}$ is uniformly bounded in the sense that

$$\sup_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}}\ \sup_{y\in[K]^n}\ \max_{t\in[n]}\|\mathbf{u}_t(y)\|_\infty<\infty.$$

This holds because $\mathbf{u}_t(y)$ is either equal to $\mathbf{v}_t(y)$, which is finite, or is equal to $f(\mathbf{x}_t(y))$ for some $f\in\mathcal{F}$, and the class $\mathcal{F}$ was already assumed to be uniformly bounded.

To bound this term we need a variant of the sequential Rademacher complexity regret bound of (Rakhlin et al., 2010), which shows that there exists a deterministic strategy for competing against any collection of trees. This is proven in the auxiliary Lemma 25 following this proof.

In particular, for each tree class $\mathcal{F}_{\mathbf{v}}$, there exists a deterministic strategy $\hat{y}_t^{\mathbf{v}}$ that guarantees the inequality

$$\sum_{t=1}^{n} \ell(\hat{y}_t^{\mathbf{v}}, y_t) - \inf_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}} \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \leq 2 \cdot \max_{\mathbf{y},\mathbf{y}'} \mathbb{E}_{\epsilon} \sup_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(\mathbf{y}_{1:t-1}(\epsilon)), \mathbf{y}'_t(\epsilon)) \right] + 2,$$

holds for every sequence, where the supremum on the right-hand-side ranges over $[K]$-valued binary trees. Futhermore, $\widehat{y}_t^{\mathbf{v}}$ is guaranteed by Lemma 25 to lie in the class $\mathcal{Z}$. We choose this strategy for the collection $\{\hat{\mathbf{y}}^{\mathbf{v}}\}$ being minimized over in (10.15). Since the regret inequality from Lemma 25 holds deterministically for all sequences y for each $\mathbf{v}$, we have that

$$(\star\star) \leq 2 \cdot \max_{\mathbf{v}\in V^{\gamma}} \max_{\mathbf{y},\mathbf{y}'} \mathbb{E}_{\epsilon} \sup_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(\mathbf{y}_{1:t-1}(\epsilon)), \mathbf{y}'_t(\epsilon)) \right] + 2.$$

For each choice of $\mathbf{v}$, $\mathbf{y}$, $\mathbf{y}'$ at the outer supremum, we define a class of real-valued trees $W_{\mathbf{v},\mathbf{y},\mathbf{y}'}$ via $\{(\mathbf{w}_t)_{t\leq n} \,:\, \mathbf{w}_t(\epsilon) := \ell(\mathbf{u}_t(\mathbf{y}(\epsilon_{1:t-1})), \mathbf{y}'_t(\epsilon)) \mid \mathbf{u} \in \mathcal{F}_{\mathbf{v}}\}$. Lemma 26 then implies

$$(\star\star) \leq 2 \max_{\mathbf{v}\in V^{\gamma}} \max_{\mathbf{y},\mathbf{y}'} \inf_{\alpha>0} \left\{ 4\alpha n + 12 \int_{\alpha}^{\mathrm{rad}_2(W_{\mathbf{v},\mathbf{y},\mathbf{y}'})} \sqrt{n \log \mathcal{N}_2(\delta, W_{\mathbf{v},\mathbf{y},\mathbf{y}'})} d\delta \right\} + 2,$$

with the real-valued covering number $\mathcal{N}_2$ and radius rad_2 defined as in Lemma 26.

We now show how to bound this covering number in terms of the covering number for $\mathcal{F}_{\mathbf{v}}$. Suppose that Z is a collection of $\mathbb{R}^K$-valued K-ary trees that form a δ-cover for $\mathcal{F}_{\mathbf{v}}$ in the sense of Definition 15. Then we have

$$\begin{aligned}
&\sup_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}} \max_{\epsilon\in\{\pm1\}^n} \inf_{\mathbf{z}\in Z} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\ell(\mathbf{u}_t(\mathbf{y}(\epsilon)), \mathbf{y}'_t(\epsilon)) - \ell(\mathbf{z}_t(\mathbf{y}(\epsilon)), \mathbf{y}'_t(\epsilon))\right)^2} \\
&\leq \sup_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}} \max_{\epsilon\in\{\pm1\}^n} \inf_{\mathbf{z}\in Z} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \max_{y'_t\in[K]} \left(\ell(\mathbf{u}_t(\mathbf{y}(\epsilon)), y'_t) - \ell(\mathbf{z}_t(\mathbf{y}(\epsilon)), y'_t)\right)^2} \\
&\leq \sup_{\mathbf{u}\in\mathcal{F}_{\mathbf{v}}} \max_{y\in[K]^n} \inf_{\mathbf{z}\in Z} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \max_{y'_t\in[K]} \left(\ell(\mathbf{u}_t(y), y'_t) - \ell(\mathbf{z}_t(y), y'_t)\right)^2} \\
&\leq \delta.
\end{aligned}$$

This implies that for any cover of $\mathcal{F}_{\mathbf{v}}$ in the sense of Definition 15 we can construct a cover for $W_{\mathbf{v},\mathbf{y},\mathbf{y}'}$ at the same scale using the construction $\{(\mathbf{w}_t)_{t\leq n} \,:\, \mathbf{w}_t(\epsilon) := \ell(\mathbf{z}_t(\mathbf{y}(\epsilon_{1:t-1})), \mathbf{y}'_t(\epsilon)) \mid \mathbf{z} \in Z\}$. Consequently, we have

$$(\star\star) \leq 2 \max_{\mathbf{v}\in V^{\gamma}} \inf_{\alpha>0}\left\{4\alpha n + 12\int_{\alpha}^{\mathrm{rad}_2(\mathcal{F}_{\mathbf{v}})} \sqrt{n\log \mathcal{N}_2(\delta, \ell\circ\mathcal{F}_{\mathbf{v}})}d\delta\right\} + 2.$$

In light of Lemma 24, this is further upper bounded by

$$\begin{aligned}(\star\star) &\leq 2\inf_{\alpha>0}\left\{4\alpha n + 12\int_{\alpha}^{\gamma} \sqrt{n\log(\mathcal{N}_2(\delta, \ell\circ\mathcal{F}, \mathbf{x})n)}d\delta\right\} + 2\\ &\leq 2\inf_{\alpha>0}\left\{4\alpha n + 12\int_{\alpha}^{\gamma} \sqrt{n\log(\mathcal{N}_2(\delta, \ell\circ\mathcal{F})n)}d\delta\right\} + 2.\end{aligned}$$

Final bound Combining $(\star)$ and $(\star\star)$, we have

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{F}) \leq \log \mathcal{N}_2(\gamma, \ell\circ\mathcal{F}) + \inf_{\gamma\geq\alpha>0}\left\{8\alpha n + 24\int_{\alpha}^{\gamma} \sqrt{n\log(\mathcal{N}_2(\delta, \ell\circ\mathcal{F})n)}d\delta\right\} + 4.$$

for any fixed γ. Optimizing over γ yields the result. $\square$

Lemma 24. Let $\mathcal{F}_{\mathbf{v}}$ be defined as in the proof of Theorem 35 for trees $\mathbf{v}$ and $\mathbf{x}$ and scale γ. Then it holds that

1. $\mathcal{N}_2(\gamma, \ell\circ\mathcal{F}_{\mathbf{v}}) \leq 1$.
2. $\mathcal{N}_2(\alpha, \ell\circ\mathcal{F}_{\mathbf{v}}) \leq n\cdot\mathcal{N}_2(\alpha, \ell\circ\mathcal{F}, \mathbf{x})$ for all $\alpha > 0$.

Proof of Lemma 24.

First claim This is essentially by construction. Recall that each element of $\mathcal{F}_v$ is of the form

$$(\mathbf{u}_{f,y})_t(y') := f(\mathbf{x}_t(y'))\mathbb{1}\{y'_1 = y_1, \ldots, y'_{t-1} = y_{t-1}\} + \mathbf{v}_t(y')\mathbb{1}\{\neg(y'_1 = y_1, \ldots, y'_{t-1} = y_{t-1})\}.$$

for some path $y \in [K]^n$ and $f \in \mathcal{F}$ for which

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n}\max_{y''_t\in[K]}(\ell(f(\mathbf{x}_t(y)), y''_t) - \ell(\mathbf{v}_t(y), y''_t))^2} \leq \gamma. \tag{10.16}$$

These properties imply that $\{\mathbf{v}\}$ is a sequential γ-cover. Indeed, using the explicit form for $\boldsymbol{u}_{f,y}$ above, it can be seen that for each path $y' \in [K]^n$, there exists some time $1 < \tau \leq n+1$ such that

$$(\mathbf{u}_{f,y})_t(y') = \begin{cases} f(\mathbf{x}_t(y')), & \text{if } t < \tau, \\ \mathbf{v}_t(y'), & \text{if } t \geq \tau. \end{cases}$$

it also holds that $y_t = y'_t$ for all $t < \tau - 1$.

Using this representation we have that for any path $y' \in [K]^n$:

$$\begin{aligned} &\sqrt{\frac{1}{n}\sum_{t=1}^{n} \max_{y''_t \in [K]} (\ell((\mathbf{u}_{f,y})_t(y'), y''_t) - \ell(\mathbf{v}_t(y'), y''_t))^2} \\ &= \sqrt{\frac{1}{n}\sum_{t=1}^{\tau-1} \max_{y''_t \in [K]} (\ell(f(\mathbf{x}_t(y'), y''_t) - \ell(\mathbf{v}_t(y'), y''_t))^2}. \end{aligned}$$

Now use that $\mathbf{x}_1, \ldots, \mathbf{x}_{\tau-1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_{\tau-1}$ only depend on $y'_1, \ldots, y'_{\tau-2}$, and that $y'_1, \ldots, y'_{\tau-2} = y_1, \ldots, y_{\tau-2}$:

$$\begin{aligned} &= \sqrt{\frac{1}{n}\sum_{t=1}^{\tau-1} \max_{y''_t \in [K]} (\ell(f(\mathbf{x}_t(y), y''_t) - \ell(\mathbf{v}_t(y), y''_t))^2} \\ &\leq \sqrt{\frac{1}{n}\sum_{t=1}^{n} \max_{y''_t \in [K]} (\ell(f(\mathbf{x}_t(y), y''_t) - \ell(\mathbf{v}_t(y), y''_t))^2} \\ &\leq \gamma. \end{aligned}$$

Second claim Let V be a cover for $\ell \circ \mathcal{F}$ on $\mathbf{x}$ of size $\mathcal{N}_2(\alpha, \ell \circ \mathcal{F}, \mathbf{x})$. Assume $|V| < \infty$ as the claim holds trivially otherwise. We will construct from V a cover $\widetilde{V}$ for $\ell \circ \mathcal{F}_{\mathbf{v}}$ with the following procedure:

- $\widetilde{V} = \emptyset$.
- For each K-ary $\mathbb{R}^K$-valued tree $\mathbf{z} \in V$ and each time $\tau \in \{2, \ldots, n+1\}$:

– Construct tree K-ary $\mathbb{R}^K$-valued tree $\mathbf{z}^{(\tau)}$ via

$$\mathbf{z}_t^{(\tau)}(y) = \mathbf{z}_t(y)\mathbb{1}\{t < \tau\} + \mathbf{v}_t(y)\mathbb{1}\{t \geq \tau\}.$$

– Add $\mathbf{z}^{(\tau)}$ to $\widetilde{V}$.

Clearly $\left|\widetilde{V}\right| \leq n \cdot |V|$. We now show that $\widetilde{V}$ is an α-cover for $\ell \circ \mathcal{F}_{\mathbf{v}}$.

Let $\mathbf{u}_{f,y}$ be an element of $\mathcal{F}_{\mathbf{v}}$ of the form described in the proof of the first claim and let $y' \in [K]^n$ be a particular path. Let τ be such that $(\mathbf{u}_{f,y})_t(y') = f(\mathbf{x}_t(y'))\mathbb{1}\{t < \tau\} + \mathbf{v}_t(y')\mathbb{1}\{t \geq \tau\}$. Let $\mathbf{z} \in V$ be α-close to f on the path y' through $\mathbf{x}$, i.e.

$$\sqrt{\frac{1}{n}\sum_{t=1}^{n} \max_{y_t'' \in [K]} (\ell(f(\mathbf{x}_t(y'), y_t'') - \ell(\mathbf{z}_t(y'), y_t''))^2} \leq \alpha.$$

Existence of such a $\mathbf{z}$ is guaranteed by the cover property of V. We will show that $\mathbf{z}^{(\tau)}$ is α-close to $\mathbf{u}_{f,y}$ on y'. Indeed, we have

$$\begin{aligned}
&\sqrt{\frac{1}{n}\sum_{t=1}^{n} \max_{y_t'' \in [K]} (\ell((\mathbf{u}_{f,y})_t(y'), y_t'') - \ell(\mathbf{z}_t^{(\tau)}(y'), y_t''))^2} \\
&= \sqrt{\frac{1}{n}\sum_{t=1}^{\tau-1} \max_{y_t'' \in [K]} (\ell(f(\mathbf{x}_t(y'), y_t'') - \ell(\mathbf{z}_t(y'), y_t''))^2 + \frac{1}{n}\sum_{t=\tau}^{n} \max_{y_t'' \in [K]} (\ell(\mathbf{v}_t(y'), y_t'') - \ell(\mathbf{v}_t(y'), y_t''))^2} \\
&= \sqrt{\frac{1}{n}\sum_{t=1}^{\tau-1} \max_{y_t'' \in [K]} (\ell(f(\mathbf{x}_t(y'), y_t'') - \ell(\mathbf{z}_t(y'), y_t''))^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{t=1}^{n} \max_{y_t'' \in [K]} (\ell(f(\mathbf{x}_t(y'), y_t'') - \ell(\mathbf{z}_t(y'), y_t''))^2} \\
&\leq \alpha.
\end{aligned}$$

Since this argument works for any $\mathbf{u}_{f,y} \in \mathcal{F}_{\mathbf{v}}$ this establishes that $\widetilde{V}$ is an α-cover of $\mathcal{F}_{\mathbf{v}}$. $\square$

The next lemma is almost the same as the sequential Rademacher complexity bound in Rakhlin et al. (2010), with the only technical difference being that the learner competes with a class of trees rather than a class of fixed functions. It is proven using the same argument as in that paper.

Lemma 25. Let U be any collection of $\mathbb{R}^K$-valued K-ary trees of depth n. Suppose that $C := \sup_{\mathbf{u}\in U} \sup_{y\in[K]^n} \max_{t\in[n]} \|\mathbf{u}_t(y)\|_\infty < \infty$. Then there exists a strategy $\hat{z}_t$ that guarantees

$$\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{\mathbf{u}\in U} \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \leq 2 \cdot \max_{\mathbf{y},\mathbf{y}'} \mathbb{E}_\epsilon \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(\mathbf{y}_{1:t-1}(\epsilon)), \mathbf{y}'_t(\epsilon)) \right] + 2,$$

where $\mathbf{y}$ and $\mathbf{y}'$ are $[K]$-valued binary trees of depth n and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are Rademacher random variables.

Furthermore, the predictions $(\hat{z}_t)_{t\leq n}$ satisfy $\|\hat{z}_t\|_\infty \leq \log(Kn)$.

Proof of Lemma 25. Define $\mathcal{Z} := \left\{ z \in \mathbb{R}^K \mid \|z\|_\infty \leq C \right\}$. The minimax optimal regret amongst deterministic strategies taking values in $\mathcal{Z}$ is given by

$$\mathcal{V}_n^{\mathrm{ol}}(U) := \left\langle\!\!\left\langle \inf_{\hat{z}_t\in\mathbb{R}^K} \max_{y_t\in[K]} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\hat{z}_t, y_t) - \inf_{\mathbf{u}\in U} \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \right].$$

Once again, this proof closely follows the sequential Rademacher complexity bound from Rakhlin et al. (2010). We only sketch the first few steps for this proof as they are identical to the first few steps of the proof of Theorem 35, which is admissible due to compactness of $\mathcal{Z}$. Using the minimax swap as in that theorem, we can move to an upper bound of

$$\begin{aligned}
&\leq \left\langle\!\!\left\langle \sup_{p_t\in\Delta_K} \mathbb{E}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \inf_{\hat{z}_t\in\mathcal{Z}} \mathbb{E}_{y_t\sim p_t} \left[\ell(\hat{z}_t, y_t) \right] - \inf_{\mathbf{u}\in U} \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \right] \\
&= \left\langle\!\!\left\langle \sup_{p_t\in\Delta_K} \mathbb{E}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \inf_{\hat{z}_t\in\mathcal{Z}} \mathbb{E}_{y_t\sim p_t} \left[\ell(\hat{z}_t, y_t) \right] - \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \right].
\end{aligned}$$

Now we choose $\hat{z}_t$ to match the value of $\mathbf{u}_t(y) = \mathbf{u}_t(y_{1:t-1})$, which is possible by definition of $\mathcal{Z}$:

$$\leq \left\langle\!\!\left\langle \sup_{p_t\in\Delta_K} \mathbb{E}_{y_t\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \mathbb{E}_{y_t\sim p_t} \left[\ell(\mathbf{u}_t(y), y_t) \right] - \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \right].$$

Using Jensen's inequality, we pull the conditional expectaitons in the first term outside the supremum over $\mathbf{u}$ by introducing a tangent sequence $(y'_t)_{t\leq n}$, where y'_t follows the distribution p_t conditioned on $y_{1:t-1}$.

$$\leq \left\langle\!\!\left\langle \sup_{p_t \in \Delta_K} \mathbb{E}_{y_t, y'_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y'_t) - \sum_{t=1}^{n} \ell(\mathbf{u}_t(y), y_t) \right].$$

Since y_t and y'_t are conditionally i.i.d., we can introduce a Rademacher random variable ϵ_t at each timestep t as follows:

$$= \left\langle\!\!\left\langle \sup_{p_t \in \Delta_K} \mathbb{E}_{y_t, y'_t \sim p_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t(\ell(\mathbf{u}_t(y), y'_t) - \ell(\mathbf{u}_t(y), y_t)) \right].$$

To decouple the arguments to the losses from the arugments to the tree $\mathbf{u}$, we move to a pessimistic upper bound:

$$\leq \left\langle\!\!\left\langle \sup_{p_t \in \Delta_K} \mathbb{E}_{y_t \sim p_t} \max_{y'_t, y''_t \in [K]} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t(\ell(\mathbf{u}_t(y), y'_t) - \ell(\mathbf{u}_t(y), y''_t)) \right]$$
$$= \left\langle\!\!\left\langle \max_{y_t, y'_t, y''_t \in [K]} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t(\ell(\mathbf{u}_t(y), y'_t) - \ell(\mathbf{u}_t(y), y''_t)) \right].$$

We now complete the symmetrization as follows:

$$\leq \left\langle\!\!\left\langle \max_{y_t, y'_t, y''_t \in [K]} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(y), y'_t) \right] + \left\langle\!\!\left\langle \max_{y_t, y'_t, y''_t \in [K]} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(y), y''_t) \right]$$
$$= 2 \cdot \left\langle\!\!\left\langle \max_{y_t, y'_t \in [K]} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(y), y'_t) \right]$$
$$= 2 \cdot \max_{\mathbf{y}, \mathbf{y}'} \mathbb{E}_{\epsilon} \sup_{\mathbf{u}\in U} \left[\sum_{t=1}^{n} \epsilon_t \ell(\mathbf{u}_t(\mathbf{y}_{1:t-1}(\epsilon)), \mathbf{y}'_t(\epsilon)) \right].$$

In the last line $\mathbf{y}$ and $\mathbf{y}'$ are taken to be $[K]$-valued binary trees of depth n, so that $\mathbf{y}_t(\epsilon) = \mathbf{y}_t(\epsilon_1, \ldots \epsilon_{t-1})$ and likewise for $\mathbf{y}'$.

Finally, to guarantee the boundedness of predictions claimed in the lemma statement, we apply Lemma 21 to the minimax optimal strategy, for which we just showed regret is bounded by the sequential Rademacher complexity. $\square$

The last auxiliary lemma in this section is a slight variant of the Dudley entropy integral bound for sequential Rademacher complexity. This lemma can be extracted from the proof of Theorem 4 in Rakhlin et al. (2015). We do not repeat the proof here.

Lemma 26. Let W be a collection of $\mathbb{R}$-valued binary trees. Define $\mathcal{N}_p(\alpha, W)$ to be the size of the smallest class of trees V such that

$$\forall \mathbf{w} \in W, \epsilon \in \{\pm 1\}^n,\ \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{n}(\mathbf{w}_t(\epsilon) - \mathbf{v}_t(\epsilon))^p\right)^{1/p} \leq \alpha. \tag{10.17}$$

Let $\mathrm{rad}_p(W) := \min\{\alpha \mid \mathcal{N}_p(\alpha, W) = 1\}$. Then it holds that

$$\mathbb{E}_\epsilon \sup_{\mathbf{w}\in W} \sum_{t=1}^{n} \epsilon_t \mathbf{w}_t(\epsilon) \leq \inf_{\alpha>0}\left\{4\alpha n + 12\int_{\alpha}^{\mathrm{rad}_2(W)} \sqrt{n \log \mathcal{N}_2(\delta, W)}d\delta\right\}. \tag{10.18}$$

10.7.4 Proof of Theorem 36

Proof of Theorem 36. First, note that an easy calculation on the softmax function $\boldsymbol{\sigma}$ implies that for all $k \in [K]$, $p_t(k) \geq \frac{(1-\mu)\exp(-2BR)+\mu}{K}$. So, defining $L = \frac{K}{(1-\mu)\exp(-2BR)+\mu}$, we have $\|\tilde{y}_t\|_1 \leq L$. Thus, Theorem 32 applied to $\mathcal{A}$ guarantees that for any $W \in \mathcal{W}$,

$$\sum_{t=1}^{n} \ell(\hat{z}_t, \tilde{y}_t) - \sum_{t=1}^{n} \ell(Wx_t, \tilde{y}_t) \leq 5LdK \cdot \log\left(\tfrac{BRn}{dK} + e\right) + 2\mu \sum_{t=1}^{n} \|\tilde{y}_t\|_1.$$

Fix a round t and let $\mathbb{E}_t[\cdot]$ denote expecation conditioned on $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_{t-1}$. The construction of the feedback vectors $\tilde{y}_t$ via importance weighting guarantees $\mathbb{E}_t[\tilde{y}_t] = \mathbb{1}_{y_t}$, where $\mathbb{1}_k$ denotes the indicator vector supported on coordinate k. Hence, $\mathbb{E}_t[\ell(\hat{z}_t, \tilde{y}_t)] = \ell(\hat{z}_t, y_t) = -\log(p_t(y_t))$ and $\mathbb{E}_t[\ell(Wx_t, \tilde{y}_t)] = \ell(Wx_t, y_t)$. Furthermore, it is easy to check that $\mathbb{E}_t[\|\tilde{y}_t\|_1] = 1$. Thus, we conclude that

$$\sum_{t=1}^{n} \mathbb{E}[-\log(p_t(y_t))] - \sum_{t=1}^{n} \ell(Wx_t, y_t) \leq 5LdK \cdot \log\left(\tfrac{BRn}{dK} + e\right) + 2\mu n.$$

Now if we set $\mu = 0$, then the right-hand side is bounded by $O(dK^2\exp(2BR)\log\left(\frac{BRn}{dK} + e\right))$. If we set $\mu = \sqrt{\frac{dK^2\log\left(\frac{BRn}{dK}+e\right)}{n}}$, the right-hand side is bounded by $O\left(\sqrt{dK^2\log(\frac{BRn}{dK} + e)n}\right)$.

Choosing the setting of μ that gives the smaller upper bound, and the fact that the log loss upper bounds the probability of making a mistake (because $-\log(p_t(y_t)) \geq 1 - p_t(y_t)$), we get the stated bound on the expected number of mistakes. $\square$

10.7.5 Proofs from Section 10.7.5

Proof of Theorem 37. Denote the number of mistakes of the i-th expert (which is the combination of the first i weak learners) by

$$M_i = \sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t^i \neq y_t\} = \sum_{t=1}^{n} \mathbb{1}\left\{\arg\max_k s_t^i(k) \neq y_t\right\},$$

with the convention that $M_0 = n$. The weights v_t^i simply implement the multiplicative weights strategy, and so Lemma 28, which gives a concentration bound based on Freedman's inequality implies that with probability at least $1-\delta$,[9]

$$\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} \leq 4 \min_i M_i + 2\log(N/\delta). \tag{10.19}$$

Note that if $k^\star := \arg\max_k s_t^{i-1}(k) \neq y_t$, then $\boldsymbol{\sigma}(s_t^{i-1})_{k^\star} \geq \boldsymbol{\sigma}(s_t^{i-1})_{y_t}$ and $\boldsymbol{\sigma}(s_t^{i-1}) \in \Delta_K$ imply $\boldsymbol{\sigma}(s_t^{i-1})_{y_t} \leq 1/2$, which then implies $\sum_{k \neq y_t} \boldsymbol{\sigma}(s_t^{i-1})_k \geq 1/2$ and finally

$$-\sum_{t=1}^{n} \widehat{C}_t^i(y_t, y_t) = \sum_{t=1}^{n} \sum_{k \neq y_t} \boldsymbol{\sigma}(s_t^{i-1})_k \geq \frac{M_{i-1}}{2}. \tag{10.20}$$

This also holds for $i = 1$ because $s_t^0 = 0$ and $-C_t^1(y_t, y_t) = (K-1)/K \geq 1/2$.

We now examine the regret guarantee provided by each logistic regression instance. For each $i \in [N]$ we have

$$\sum_{t=1}^{n} \ell(s_t^i, y_t) - \inf_{W \in \mathcal{W}} \sum_{t=1}^{n} \ell(W\widetilde{x}_t^i, y_t) \leq O(\log(n \log(nK)))$$

[9]Note that previous online boosting works (Beygelzimer et al., 2015; Jung et al., 2017) use a simpler Hoeffding bound at this stage, which picks up an extra $\sqrt{n}$ term. For their results this is not a dominant term, but in our case it can spoil the improvement given by improper logistic regression, and so we use Freedman's inequality to remove it.

This follows from Theorem 32 using $L = 1$, $D_{\mathcal{W}} = 1$, $B = 3$ for ℓ_1 norm, $\|y_t\|_1 = 1$, $\mu = 1/n$, and $\|\widetilde{x}^i_t\|_\infty \leq \log(nK)$, where the last fact is implied by the second statement of Theorem 32: $\|s^i_t\|_\infty \leq \log(K/\mu) = \log(nK)$ and thus $\|\widetilde{x}^i_t\|_\infty = \left\|(e_{l^i_t}, s^{i-1}_t)\right\|_\infty \leq \log(nK)$. Now define the difference between the total loss of the i-th and $(i-1)$-th expert to be

$$\Delta_i = \sum_{t=1}^{n} \ell(s^i_t, y_t) - \ell(s^{i-1}_t, y_t).$$

Since $\inf_{W \in \mathcal{W}} \sum_{t=1}^n \ell(W\widetilde{x}^i_t, y_t) = \inf_{\alpha \in [-2,2]} \sum_{t=1}^n \ell(\alpha e_{l^i_t} + s^{i-1}_t, y_t)$, the regret bound above implies

$$\Delta_i \leq \inf_{\alpha \in [-2,2]} \left[\sum_{t=1}^{n} \ell(\alpha e_{l^i_t} + s^{i-1}_t, y_t) - \ell(s^{i-1}_t, y_t) \right] + O(\log(n \log(nK))).$$

By Lemma 29 each term in the sum above satisfies

$$\ell(\alpha e_{l^i_t} + s^{i-1}_t, y_t) - \ell(s^{i-1}_t, y_t) \leq \begin{cases} (e^\alpha - 1)\boldsymbol{\sigma}(s^{i-1}_t)_{l^i_t} = (e^\alpha - 1)\widehat{C}^i_t(y_t, l^i_t), & l^i_t \neq y_t, \\ (e^{-\alpha} - 1)(1 - \boldsymbol{\sigma}(s^{i-1}_t)_{y_t}) = -(e^{-\alpha} - 1)\widehat{C}^i_t(y_t, y_t), & l^i_t = y_t. \end{cases}$$

With notation $w^i = -\sum_{t=1}^n \widehat{C}^i_t(y_t, y_t)$, $c^i_+ = -\frac{1}{w^i} \sum_{t: l^i_t = y_t} \widehat{C}^i_t(y_t, y_t)$, and $c^i_- = \frac{1}{w^i} \sum_{t: l^i_t \neq y_t} \widehat{C}^i_t(y_t, l^i_t)$, we rewrite

$$\inf_{\alpha \in [-2,2]} \left[\sum_{t=1}^{n} \ell(\alpha e_{l^i_t} + s^{i-1}_t, y_t) - \ell(s^{i-1}_t, y_t) \right] = w^i \cdot \inf_{\alpha \in [-2,2]} \left[(e^\alpha - 1)c^i_- + (e^{-\alpha} - 1)c^i_+ \right].$$

One can verify that $w^i > 0$, $c^i_-, c^i_+ \geq 0$, $c^i_+ - c^i_- = \gamma_i \in [-1, 1]$ and $c^i_+ + c^i_- \leq 1$. By Lemma 30, it follows that

$$w^i \cdot \inf_{\alpha \in [-2,2]} \left[(e^{-\alpha} - 1)c^i_- + (e^\alpha - 1)c^i_+ \right] \leq -\frac{w^i \gamma_i^2}{2}.$$

Summing Δ_i over $i \in [N]$, we have

$$\sum_{t=1}^{n} \ell(s^N_t, y_t) - \sum_{t=1}^{n} \ell(s^0_t, y_t) = \sum_{i=1}^{N} \Delta_i \leq -\frac{1}{2} \sum_{i=1}^{N} w^i \gamma_i^2 + O(N \log(n \log(nK))). \tag{10.21}$$

We lower bound the left hand side as

$$\sum_{t=1}^{n} \ell(s^N_t, y_t) - \sum_{t=1}^{n} \ell(s^0_t, y_t) \geq -\sum_{t=1}^{n} \ell(s^0_t, y_t) = -n \log(K),$$

where the inequality uses non-negativity of the logistic loss and the equality is a direct calculation from $s_t^0 = 0$. Next we upper bound the right-hand side of (10.21). Since $w^i = -\sum_{t=1}^n \widehat{C}_t^i(y_t, y_t)$, Eq. (10.20) implies

$$-\frac{1}{2}\sum_{i=1}^N w^i \gamma_i^2 \le -\frac{1}{4}\sum_{i=1}^N M_{i-1}\gamma_i^2 \le -\min_{i\in[N]} M_{i-1} \cdot \frac{1}{4}\sum_{i=1}^N \gamma_i^2 \le -\min_{i\in[N]} M_i \cdot \frac{1}{4}\sum_{i=1}^N \gamma_i^2.$$

Combining our upper and lower bounds on $\sum_{i=1}^N \Delta_i$ now gives

$$-n\log(K) \le -\frac{1}{2}\sum_{i=1}^N w^i\gamma_i^2 + O(N\log(n\log(K))) \le -\min_{i\in[N]} M_i \cdot \frac{1}{4}\sum_{i=1}^N \gamma_i^2 + O(N\log(n\log(nK))). \tag{10.22}$$

Rearranging, we have

$$\min_{i\in[N]} M_i \le O\left(\frac{n\log(K)}{\sum_{i=1}^N \gamma_i^2}\right) + O\left(\frac{N\log(n\log(nK))}{\sum_{i=1}^N \gamma_i^2}\right).$$

Returning to (10.19), this implies that with probability at least $1-\delta$,

$$\sum_{t=1}^n \mathbb{1}\{\widehat{y}_t \ne y_t\} \le O\left(\frac{n\log(K)}{\sum_{i=1}^N \gamma_i^2}\right) + O\left(\frac{N\log(n\log(nK))}{\sum_{i=1}^N \gamma_i^2}\right) + 2\log(N/\delta),$$

which finishes the proof. □

Proof of Proposition 18. By the definition of the cost matrices, the weak learning condition

$$\sum_{t=1}^n C_t^i(y_t, l_t^i) \le \sum_{t=1}^n \mathop{\mathbb{E}}_{k\sim u_{\gamma,y_t}}\left[C_t^i(y_t, k)\right] + S$$

implies

$$\sum_{t=1}^n \widehat{C}_t^i(y_t, l_t^i) \le \sum_{t=1}^n \mathop{\mathbb{E}}_{k\sim u_{\gamma,y_t}}\left[\widehat{C}_t^i(y_t, k)\right] + KS$$

Expanding the definitions of u_{γ,y_t} and $\widehat{C}_t^i$, we have

$$\mathop{\mathbb{E}}_{k\sim u_{\gamma,y_t}}\left[\widehat{C}_t^i(y_t, k)\right] = \left(\frac{1-\gamma}{K}\right)\left((\boldsymbol{\sigma}(s_t^{i-1})_{y_t} - 1) + \sum_{k\ne y_t}\boldsymbol{\sigma}(s_t^{i-1})_k\right) + \gamma(\boldsymbol{\sigma}(s_t^{i-1})_{y_t} - 1) = \gamma\widehat{C}_t^i(y_t, y_t).$$

So we have

$$\sum_{t=1}^{n} \widehat{C}_t^i(y_t, l_t^i) \le \gamma \sum_{t=1}^{n} \widehat{C}_t^i(y_t, y_t) + KS,$$

or, since $\widehat{C}_t^i(y_t, y_t) < 0$,

$$\gamma_i \ge \gamma - \frac{KS}{w^i},$$

where $w^i = -\sum_{t=1}^{n} C_t^i(y_t, y_t)$ as in the proof of Theorem 37. Since $a \ge b - c$ implies $a^2 \ge b^2 - 2bc$ for non-negative a, b and c, we further have $\gamma_i^2 \ge \gamma^2 - 2\frac{\gamma KS}{w^i}$.

Returning to the inequality (10.22), the bound we just proved implies

$$\begin{aligned}
-n\log(K) &\le -\frac{1}{2}\sum_{i=1}^{N} w^i\gamma^2 + \gamma KSN + O(N\log(n\log(nK))) \\
&\le -\frac{\gamma^2}{4}\sum_{i=1}^{N} M_{i-1} + \gamma KSN + O(N\log(n\log(nK))) \qquad \text{(by (10.20))} \\
&\le -\min_{i\in[N]} M_i \cdot \frac{\gamma^2 N}{4} + \gamma KSN + O(N\log(n\log(nK))).
\end{aligned}$$

From here we proceed as in the proof of Theorem 37 to get the result. □

Lemma 27 (Freedman's Inequality (Beygelzimer et al., 2011))**.** Let $(Z_t)_{t\le n}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathcal{J}_t)_{t\le n}$ with $|Z_t| \le R$ almost surely. For any $\eta \in [0, 1/R]$, with probability at least $1-\delta$,

$$\sum_{t=1}^{n} Z_t \le \eta(e-2)\sum_{t=1}^{n} \mathbb{E}\left[Z_t^2 \mid \mathcal{J}_t\right] + \frac{\log(1/\delta)}{\eta} \tag{10.23}$$

for all $\eta \in [0, 1/R]$.

Lemma 28. With probability at least $1-\delta$, the predictions $(\widehat{y}_t)_{t\le n}$ generated by Algorithm 11 satisfy

$$\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \ne y_t\} \le 4\min_i \sum_{t=1}^{n} \mathbb{1}\left\{\widehat{y}_t^i \ne y_t\right\} + 2\log(N/\delta).$$

Proof. Define a filtration $(\mathcal{J}_t)_{t\le n}$ via

$$\mathcal{J}_t = \sigma((x_1, (l_1^i)_{i\le N}, y_1, i_1), \ldots, (x_{t-1}, (l_{t-1}^i)_{i\le N}, y_{t-1}, i_{t-1}), x_t, (l_t^i)_{i\le N}).$$

Since Line 18 of Algorithm 11 implements the multiplicative weights strategy with learning rate 1, the standard analysis (e.g. Cesa-Bianchi and Lugosi (2006)) implies that the conditional expectations under this strategy enjoy a regret bound of

$$\sum_{t=1}^{n} \mathbb{E}[\mathbb{1}\{\widehat{y}_t \neq y_t\} \mid \mathcal{J}_t] \leq 2 \min_i \sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t^i \neq y_t\} + \log(N).$$

Let $Z_t = \mathbb{1}\{\widehat{y}_t \neq y_t\} - \mathbb{E}[\mathbb{1}\{\widehat{y}_t \neq y_t\} \mid \mathcal{J}_t]$. Lemma 27 applied with $\eta = 1$ shows that with probability at least $1-\delta$,

$$\sum_{t=1}^{n} Z_t \leq \sum_{t=1}^{n} \mathbb{E}\big[Z_t^2 \mid \mathcal{J}_t\big] + \log(1/\delta).$$

Since variance is bounded by second moment, we have

$$\sum_{t=1}^{n} \mathbb{E}\big[Z_t^2 \mid \mathcal{J}_t\big] \leq \sum_{t=1}^{n} \mathbb{E}\big[(\mathbb{1}\{\widehat{y}_t \neq y_t\})^2 \mid \mathcal{J}_t\big] = \sum_{t=1}^{n} \mathbb{E}[\mathbb{1}\{\widehat{y}_t \neq y_t\} \mid \mathcal{J}_t].$$

Rearranging, we have proved that with probability $1-\delta$,

$$\sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t \neq y_t\} \leq 2\sum_{t=1}^{n} \mathbb{E}[\mathbb{1}\{\widehat{y}_t \neq y_t\} \mid \mathcal{J}_t] + \log(1/\delta) \leq 4 \min_i \sum_{t=1}^{n} \mathbb{1}\{\widehat{y}_t^i \neq y_t\} + 2\log(N/\delta).$$

□

Lemma 29. The multiclass logistic loss satisfies for any $z \in \mathbb{R}^K$ and $y \in [K]$,

$$\ell(z + \alpha e_l, y) - \ell(z, y) \leq \begin{cases} (e^{\alpha} - 1)\boldsymbol{\sigma}(z)_l, & l \neq y, \\ (e^{-\alpha} - 1)(1 - \boldsymbol{\sigma}(z)_y), & l = y. \end{cases}$$

Proof. When $l \neq y$ we have

$$\begin{aligned} \ell(z + \alpha e_l, y) - \ell(z, y) &= \log\left(\frac{1 + \sum_{k \neq y, l} e^{z_k - z_y} + e^{z_l + \alpha - z_y}}{1 + \sum_{k \neq y} e^{z_k - z_y}}\right) \\ &= \log\left(1 + (e^{\alpha} - 1)\frac{e^{z_l - z_y}}{1 + \sum_{k \neq y} e^{z_k - z_y}}\right) \\ &= \log(1 + (e^{\alpha} - 1)\boldsymbol{\sigma}(z)_l) \\ &\leq (e^{\alpha} - 1)\boldsymbol{\sigma}(z)_l. \qquad (\log(1+x) \leq x) \end{aligned}$$

When $l = y$ we have

$$\begin{aligned}
\ell(z + \alpha e_l, y) - \ell(z, y) &= \log\left(\frac{1 + e^{-\alpha}\sum_{k\neq y} e^{z_k - z_y}}{1 + \sum_{k\neq y} e^{z_k - z_y}}\right) \\
&= \log\left(1 + (e^{-\alpha} - 1)\frac{\sum_{k\neq y} e^{z_k - z_y}}{1 + \sum_{k\neq y} e^{z_k - z_y}}\right) \\
&= \log\left(1 + (e^{-\alpha} - 1)\sum_{k\neq y} \boldsymbol{\sigma}(z)_k\right) \\
&= \log\big(1 + (e^{-\alpha} - 1)(1 - \boldsymbol{\sigma}(z)_y)\big) \\
&\leq (e^{-\alpha} - 1)(1 - \boldsymbol{\sigma}(z)_y). \qquad\qquad (\log(1+x) \leq x)
\end{aligned}$$

□

Lemma 30 (Jung et al. (2017))**.** For any $A, B \geq 0$ with $A - B \in [-1, +1]$ and $A + B \leq 1$,

$$\inf_{\alpha\in[-2,2]} \big[A(e^{\alpha} - 1) + B(e^{-\alpha} - 1)\big] \leq -\frac{(A - B)^2}{2}.$$

10.7.6 Efficient Implementation

In this section we discuss an efficient (i.e. polynomial time in the parameters of the problem) randomized implementation of Algorithm 9. The main idea is to exploit the log-concavity of the density of P_t in the algorithm and to use well-established Markov chain Monte Carlo samplers for such densities to collect enough matrices W sampled from the distribution to approximate the prediction $\hat{z}_t$ sufficiently well to ensure the increase in regret is small.

Fix a round t. Recall that the density on $\mathcal{W}$ we wish to sample from in round t of the algorithm is

$$P_t(W) \propto \exp(-\tfrac{1}{L}\textstyle\sum_{s=1}^{t-1}\ell(W x_s, y_s)).$$

For notational convenience, define the function $F_t : \mathcal{W} \to \mathbb{R}$ as $F_t(W) := \exp(-\frac{1}{L}\sum_{s=1}^{t-1}\ell(Wx_s, y_s))$. It is easy to check that F_t is log-concave.

Assumption 6. We have access to a sampler that makes $\mathrm{poly}(1/\varepsilon, n, d, B, R)$ queries to F_t and produces a sample W with distribution $\tilde{P}_t$ such that $d_{\mathrm{TV}}(\tilde{P}_t, P_t) \leq \varepsilon$.

Such samplers are well-known in the literature: for example, the hit-and-run sampler (Lovász and Vempala, 2006), the projected Langevin Monte Carlo sampler (Bubeck et al., 2018), and the Dikin walk sampler (Narayanan and Rakhlin, 2017). It is easy to derive appropriate bounds on all the relevant parameters of F_t that are involved in the analysis of these samplers so that the samplers run in polynomial time. While this gives a theoretically efficient implementation, the running time bounds are too loose to be practical (for example, see the calculations below for projected Langevin Monte Carlo sampler). We have not attempted to improve these running time bounds; that is a direction for future work.

Example 23 (Bubeck et al. (2018))**.** *Let W have density $P \propto e^{-f}$ for some β-smooth, S-Lipschitz convex function f over a convex body $\mathcal{W}$ contained in a euclidian ball of radius D in dimension d. Projected Langevin Monte Carlo produces a sample from $\widetilde{P}$ with $d_{TV}(\widetilde{P}, P) \leq \varepsilon$ after $O\left(\frac{D^6 \max\{d, DS, D\beta\}^{12}}{\varepsilon^{12}}\right)$ evaluations. For our setting, when $\|x_t\|_2 \leq R$ and $\|y_t\|_1 \leq L$, the loss $w \mapsto \ell(\langle w, x_t\rangle, y_t)$ is $O(RL)$-Lipschitz and smooth. We therefore have $S, \beta \leq RLn$ and $D = B$, which yields the following bound on the number of queries to F_t:*

$$O\left(\frac{B^6 \max\{dK, BRLn\}^{12}}{\varepsilon^{12}}\right).$$

Given access to a sampler, we can now prove Proposition 17. In the following, we use the phrase "with high probability" to indicate that the statement referred to holds with probability at least $1 - \delta$ for any $\delta > 0$. We also use the notation $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to suppress logarithmic dependence on $1/\delta$, d, K, and n.

Proof of Proposition 17. The idea is very straightforward: for some parameters $m \in \mathbb{N}$ and $\varepsilon > 0$ to be specified later, in each round t, simply use the sampler with error tolerance $\frac{\varepsilon}{2}$ repeatedly m times to collect samples $W^{(i)}$ for $i \in [m]$ and then approximate the prediction by $\tilde{z}_t = \boldsymbol{\sigma}^+\big(\mathrm{smooth}_\mu\big(\mathbb{E}_{i\sim[m]}\big[\boldsymbol{\sigma}(W^{(i)}x_t)\big]\big)\big)$. Here, "$i \sim [m]$" denotes sampling i uniformly from $[m]$, and $m = \mathrm{poly}(n, d, B, R, 1/\delta)$ will be chosen to be large enough to ensure that this approximation incurs only $1/n$ additional loss in each round, with high probability, and thus at most $O(1)$ additional loss over all n rounds.

It remains to provide appropriate bounds on m. In the following, we will fix the round t and drop the subscript t from $P_t, \tilde{P}_t, x_t, y_t$, etc. for notational clarity.

Define the distributions $p = \mathrm{smooth}_\mu(\mathbb{E}_{W\sim P}[\boldsymbol{\sigma}(Wx)])$, $\tilde{p} = \mathrm{smooth}_\mu(\mathbb{E}_{W\sim \tilde{P}}[\boldsymbol{\sigma}(Wx)])$ and $\tilde{\tilde{p}} = \mathrm{smooth}_\mu\big(\mathbb{E}_{i\sim[m]}\big[\boldsymbol{\sigma}(W^{(i)}x)\big]\big)$. Then standard Chernoff-Hoeffding bounds and a union bound over all $k \in [K]$ imply that if $m = \tilde{\Omega}(1/\varepsilon^2)$, then with high probability, we have $\|\tilde{p} - \tilde{\tilde{p}}\|_\infty \le \frac{\varepsilon}{2}$. Furthermore, the sampler ensures $d_{\mathrm{TV}}(\tilde{P}, P) \le \frac{\varepsilon}{2}$, which implies that $\|p - \tilde{p}\|_\infty \le \frac{\varepsilon}{2}$ since each coordinate of p and $\tilde{p}$ are i n $[0, 1$. Thus, by the triangle inequality, we have $\|p - \tilde{\tilde{p}}\|_\infty \le \varepsilon$.

We now bound the excess loss for using $\tilde{\tilde{p}}$ instead of p in the algorithm, using the fact the weighted multiclass logistic loss can be equivalently viewed as a weighted multiclass log loss after passing the logits through the softmax function $\boldsymbol{\sigma}$. Thus, the additional loss equals

$$\sum_{k\in[K]} y_k \log\Big(\frac{p_k}{\tilde{\tilde{p}}_k}\Big) \le \sum_{k\in[K]} y_k \log\Big(\frac{\tilde{\tilde{p}}_k+\varepsilon}{\tilde{\tilde{p}}_k}\Big) \le \sum_{k\in[K]} y_k \log\Big(1 + \frac{\varepsilon K}{\mu}\Big) \le \frac{\varepsilon K L}{\mu}.$$

The first inequality above follows from the bound $\|p - \tilde{\tilde{p}}\|_\infty \le \varepsilon$, and the second from the fact that $\tilde{\tilde{p}}_k \ge \frac{\mu}{K}$ for all $k \in [K]$, and the third from $\log(1 + a) \le a$ for all $a \in \mathbb{R}_+$ and $\|y\|_1 \le L$. Thus, setting $\varepsilon = \frac{\mu}{KLn}$ ensures that the additional loss is at most $1/n$ with high probability, as required. □

10.8 Chapter Notes

This chapter is based on Foster et al. (2018b).

CHAPTER 11

CONTEXTUAL BANDITS

In this chapter we develop adaptive learning guarantees for the contextual bandit setting. The contextual bandit setting is a sequential decision making model that generalizes the online supervised learning to accommodate partial feedback, and has been successfully applied in content recommendation and beyond (Li et al., 2010; Agarwal et al., 2016; Tewari and Murphy, 2017; Greenewald et al., 2017).

We introduce a new family of margin-based regret guarantees for adversarial contextual bandit learning. The new margin bound serves as a generic contextual bandit analogue of the classical margin bound from statistical learning. This result is based on multiclass surrogate losses, combined with the minimax analysis techniques for adaptive online learning developed in Part II. Using the ramp loss, we derive a generic margin-based regret bound in terms of the sequential metric entropy for a benchmark class of real-valued regression functions. The result applies to large nonparametric classes, improving on the best known results for Lipschitz contextual bandits (Cesa-Bianchi et al., 2017) and, as a special case, generalizes the dimension-independent BANDITRON regret bound (Kakade et al., 2008) to arbitrary linear classes with smooth norms. Under realizability assumptions our results also yield classical regret bounds.

On the algorithmic side, we use the hinge loss to derive an efficient algorithm with a $\sqrt{dn}$-type mistake bound against benchmark policies induced by d-dimensional regression functions. This provides the first hinge loss-based solution to the open problem of Abernethy and Rakhlin (2009). With an additional i.i.d. assumption, we give a simple oracle-efficient algorithm whose regret matches our generic metric entropy-based bound for sufficiently complex nonparametric classes.

11.1 Background

Surrogate loss functions are ubiquitous in supervised learning (cf. Zhang (2004); Bartlett et al. (2006); Schapire and Freund (2012)). Computationally, they are used to replace NP-hard optimization problems with computationally tractable ones, e.g., the hinge loss makes binary classification amenable to convex programming techniques. Statistically, they also enable sharper generalization analysis for models including boosting, SVMs and neural networks (Schapire and Freund, 2012; Anthony and Bartlett, 2009), for example by replacing dependence on dimension in VC-type bounds with distribution-dependent quantities.

In this chapter, we use surrogate loss functions to derive a new family of *margin-based* algorithms and regret bounds for contextual bandits. Curiously, surrogate losses have seen limited use in partial information settings (some exceptions are discussed below). This chapter demonstrates that these desirable computational and statistical properties indeed extend to contextual bandits.

In the first part of the chapter we focus on statistical issues, namely whether *any algorithm* can achieve a generalization of the classical margin bound from statistical learning (Boucheron et al., 2005) in the adversarial contextual bandit setting. Our aim here is to introduce a generic margin-based guarantees, in analogy with statistical and online learning, and our results provide an information-theoretic benchmark for future algorithm designers. We consider benchmark policies induced by a class $\mathcal{F}$ of real-valued regression functions, and the achievability results we present depend on the complexity of $\mathcal{F}$. As one consequence, we show that $\tilde{O}(n^{\frac{d}{d+1}})$ regret is achievable for Lipschitz contextual bandits in

d-dimensional metric spaces, improving on a recent result of Cesa-Bianchi et al. (2017), and that an $\tilde{O}(n^{2/3})$ mistake bound is achievable for bandit multiclass prediction in smooth Banach spaces (extending Kakade et al. (2008)).

Technically, to provide an analogue of the classical margin theory, we must overcome several challenges. First, since we operate in the online adversarial setting, there is no generic algorithmic counterpart to empirical risk minimization that we can use to analyze statistical behavior of arbitrary classes. Instead, we build on the non-constructive adaptive minimax analysis developed in Part II, specifically Chapter 6. Since we work in the contextual bandit setting, we must extend these arguments to incorporate partial information.

In the second part of the chapter, we focus on computational issues and derive two new algorithms using the hinge loss as a convex surrogate. The first algorithm, HINGE-LMC, provably runs in polynomial time and achieves a $\sqrt{dn}$-type mistake bound against d-dimensional benchmark regressors with suitable convexity properties. HINGE-LMC is the first efficient algorithm with $\sqrt{dn}$-mistake bound for bandit multiclass prediction using a surrogate loss without curvature, and so it provides a new resolution to the open problem of Abernethy and Rakhlin (2009). This algorithm is based on the exponential weights update, along with Langevin Monte Carlo for efficient sampling and a careful action selection scheme to ensure low regret. The second algorithm is much simpler: we show that, in the stochastic setting, Follow-The-Leader with appropriate smoothing provides an algorithmic counterpart to the aforementioned information-theoretic results provided the class $\mathcal{F}$ is sufficiently large (in terms of metric entropy growth rate). We caution that compared to preceding chapters, the algorithmic techniques we employ are somewhat ad-hoc and do not exactly fall into the Burkholder framework. Understanding the extent to which the Burkholder method can generically be used to design contextual bandit algorithms is an important direction for future research.

11.1.1 Preliminaries

We work in the contextual bandit protocol (Protocol 4), with loss space $\mathcal{L} = [0,1]^{\mathcal{A}}$. Recall that the goal is to design algorithms that achieve low *regret* against a class $\Pi \subset (\mathcal{X} \to \mathcal{A})$ of benchmark policies:

$$\mathrm{Reg}_n(n, \Pi) \triangleq \sum_{t=1}^{n} \mathbb{E}[\ell_t(a_t)] - \inf_{\pi \in \Pi} \sum_{t=1}^{n} \mathbb{E}[\ell_t(\pi(x_t))].$$

In this chapter, we always identify Π with a class of vector-valued regression functions $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R}^K_{=0})$, where we define $\mathbb{R}^K_{=0} \triangleq \{s \in \mathbb{R}^K : \sum_a s_a = 0\}$. For such functions, we use the notation $f(x) \in \mathbb{R}^K$ to denote the vector-valued output and $f(x)_a$ to denote the a^{th} component. Note that we are assuming $\sum_a f(x)_a = 0$, which is a natural generalization of the standard regression function formulation of binary classification (Bartlett et al., 2006) and appears in e.g. Pires et al. (2013). We define $B \triangleq \sup_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} \|f(x)\|_\infty$ to be the maximum value predicted by any regressor.

Our algorithms use *importance weighting* to form unbiased loss estimates. If at round t, the algorithm chooses action a_t by sampling from a distribution $p_t \in \Delta([K])$, the loss estimate is defined as $\hat{\ell}_t(a) \triangleq \ell_t(a_t)\mathbf{1}\{a_t = a\}/p_t(a)$. Given p_t, we also define a smoothed distribution as $p_t^\mu \triangleq (1 - K\mu)p_t + \mu$ for some parameter $\mu \in [0, 1/K]$.

We introduce two surrogate loss functions, the *ramp loss* and the *hinge loss*, whose scalar versions are defined as $\phi^\gamma(s) \triangleq \min(\max(1 + s/\gamma, 0), 1)$ and $\psi^\gamma(s) \triangleq \max(1 + s/\gamma, 0)$ respectively. For $s \in \mathbb{R}^K$, $\phi^\gamma(s)$ and $\psi^\gamma(s)$ are defined coordinate-wise.

We start with a simple lemma, demonstrating how ϕ^γ, ψ^γ act as surrogates for cost-sensitive multiclass losses.

Lemma 31 (Surrogate Loss Translation)**.** For $s \in \mathbb{R}^K_{=0}$, define $\pi_{\text{ramp}}(s) \in \Delta(\mathcal{A})$ by $\pi_{\text{ramp}}(s)_a \propto \phi^\gamma(s_a)$ and define $\pi_{\text{hinge}}(s) \in \Delta(\mathcal{A})$ by $\pi_{\text{hinge}}(s)_a \propto \psi^\gamma(s_a)$ analogously. For any vector $\ell \in \mathbb{R}^K_+$, we have

$$\langle \pi_{\text{ramp}}(s), \ell \rangle \leq \langle \ell, \phi^\gamma(s) \rangle \leq \sum_{a \in \mathcal{A}} \ell(a) \mathbf{1}\{s_a \geq -\gamma\}, \qquad \text{and} \qquad \langle \pi_{\text{hinge}}(s), \ell \rangle \leq K^{-1} \langle \ell, \psi^\gamma(s) \rangle.$$

Based on this lemma, it will be convenient to define $L_n^\gamma(f) \triangleq \sum_{t=1}^n \sum_{a \in \mathcal{A}} \ell_t(a) \mathbf{1}\{f(x_t)_a \geq -\gamma\}$, which is the *margin-based cumulative loss* for the regressor f. L_n^γ should be seen as a cost-sensitive multiclass analogue of the classical margin loss from statistical learning (Boucheron et al., 2005).

11.2 Minimax Achievability of Margin Bounds

This section provides generic margin-based regret bounds for contextual bandits in terms of the sequential metric entropy of the regressor class $\mathcal{F}$. Notably, our general techniques apply when the ramp loss is used as a surrogate, and so they yield the main result of the section—a margin-based regret guarantee—as a special case.

To motivate our approach, consider a well-known reduction from bandits to full information online learning: If a full information algorithm achieves a regret bound in terms of the so-called *local norms* $\sum_t \langle p_t, \ell_t^2 \rangle$, then running the full information algorithm on importance-weighted losses $\hat{\ell}_t(a)$ yields an expected regret bound for the bandit setting. For example, EXP4 (Auer et al., 2002b) exploits this observation for the case when Π is finite, using HEDGE (Freund and Schapire, 1997) as the full information algorithm, and obtaining a deterministic regret bound of

$$\mathrm{Reg}_n(n, \Pi) \leq \frac{\eta}{2} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\pi \sim p_t} \left\langle \pi(x_t), \hat{\ell}_t \right\rangle^2 + \frac{\log(|\Pi|)}{\eta}, \tag{11.1}$$

where $\eta > 0$ is the learning rate and p_t is the distribution over policies in Π (which induces an action distribution for round t). Evaluating conditional expectations and optimizing the learning rate η yields a regret bound of $\mathcal{O}(\sqrt{Kn \log(|\Pi|)})$, which is optimal for contextual bandits with a finite policy class.

To use this reduction beyond the finite class case and with surrogate losses we face two challenges:

1. **Infinite Classes.** The natural approach of using a pointwise (or sup-norm) cover of the function class $\mathcal{F}$ is insufficient—not only because there are classes that have infinite pointwise covers yet are online-learnable, but also because it yields sub-optimal rates even when a finite pointwise cover is available. Instead, we directly establish existence of a full-information algorithm for large, potentially nonparametric classes that has 1) strong adaptivity to loss scaling similar to (11.1) and 2) regret scaling with the sequential covering number for $\mathcal{F}$, which is the correct generalization of the empirical covering number in statistical learning to the adversarial online setting. This is achieved by using the tools of Part II to establish achievability.
2. **Variance Control**. With surrogate losses, controlling the variance term $\mathbb{E}_{\pi}\langle \pi(x_t), \ell_t \rangle^2$ in the reduction from bandit to full information is more challenging, since the surrogate loss of a policy depends on the scale of the underlying regressor, not just the action it selects. To address this, we develop a new sampling scheme tailored to scale-sensitive losses.

Full-Information Regret Bound. We consider the following full information protocol, which in the sequel will be instantiated via reduction from contextual bandits. Let the context space $\mathcal{X}$ and $\mathcal{A}$ be fixed as in Section 11.1.1, and consider a function class $\mathcal{G} \subset (\mathcal{X} \to \mathcal{S})$, where $\mathcal{S} \subseteq \mathbb{R}_+^K$. The reader may think of $\mathcal{G}$ as representing $\phi^\gamma \circ \mathcal{F}$ or $\psi^\gamma \circ \mathcal{F}$, i.e. the surrogate loss composed with the regressor class, so that $\mathcal{S}$ (which is not necessarily convex) represents the image of the surrogate loss over $\mathcal{F}$.

The online learning protocol is: For time $t = 1, \ldots, n$, (1) the learner picks a distribution $p_t \in \Delta(\mathcal{S})$, (2) the adversary picks a loss vector $\ell_t \in \mathcal{L} \subset \mathbb{R}_+^K$, (3) the learner samples outcome $s_t \sim p_t$ and experiences loss $\langle s_t, \ell_t \rangle$. Regret against the benchmark class $\mathcal{G}$ is given by

$$\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s_t, \ell_t \rangle - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle.$$

Similar to Chapter 10, we require a multi-output generalization of the *sequential covering numbers* described in Chapter 6 because we work in the multiclass setting.

Definition 17. *For a function class $\mathcal{G} : \mathcal{X} \to \mathbb{R}^K$ and $\mathcal{X}$-valued tree $\boldsymbol{x}$ of length n, the L_∞/ℓ_∞ sequential covering number for $\mathcal{G}$ on $\boldsymbol{x}$ at scale ε, denoted by $\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, \boldsymbol{x})$, is the cardinality of the smallest set V of $\mathbb{R}^K$-valued trees for which*

$$\forall g \in \mathcal{G}\ \forall \epsilon \in \{\pm 1\}^n\ \exists \boldsymbol{v} \in V \text{ s.t. } \max_{t \in [n]} \| g(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{v}_t(\epsilon) \|_\infty \leq \varepsilon.^{1} \tag{11.2}$$

Define $\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n) = \sup_{\boldsymbol{x}:\mathrm{length}(\boldsymbol{x})=n} \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, \boldsymbol{x})$.

We refer to $\log \mathcal{N}_{\infty,\infty}$ as the *sequential metric entropy*. Note that in the binary case, for learning unit ℓ_2 norm linear functions in d dimensions, the pointwise metric entropy grows as $O(d \log(1/\varepsilon))$, whereas the sequential metric entropy is $O(d \log(1/\varepsilon) \wedge \varepsilon^{-2} \log(d))$, leading to improved rates in high dimension.

[1] Sequential coverings for L_p/ℓ_q can be defined similarly, but do not appear in the present chapter.

With this definition, we can now state our main theorem for full information.

Theorem 40. *Assume* $\sup_{\ell \in \mathcal{L}} \|\ell\|_1 \leq R$[2] *and* $\sup_{s \in \mathcal{S}} \|s\|_\infty \leq B$. *Fix any constants* $\eta \in (0,1]$, $\lambda > 0$, *and* $\beta > \alpha > 0$. *Then there exists an algorithm with the following deterministic regret guarantee:*

$$\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s, \ell_t \rangle - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle \leq \frac{2\eta}{RB} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s_t, \ell_t \rangle^2 + \frac{4RB}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) + 3e^2 \alpha \sum_{t=1}^{n} \|\ell_t\|_1$$
$$+ 12e \left(\frac{\lambda}{4R} \sum_{t=1}^{n} \|\ell_t\|_1^2 + \frac{R}{\lambda} \right) \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon.$$

Observe that the bound involves the variance terms/local norms $\mathbb{E}_{s_t \sim p_t} \langle s_t, \ell_t \rangle^2$, and has a very mild explicit dependence on the loss range R; this can be verified by optimizing over η and λ. This adaptivity to the loss range is crucial for our bandit reduction. Further observe that the bound contains a Dudley-type entropy integral, which is essential for obtaining sharp rates for complex nonparametric classes. The proof of Theorem 40 follows similar reasoning to the achievability results in Chapter 6, particularly the Online PAC-Bayes theorem. It is substantially more technical because a) the regret bound depends on the learner's own predictions and so does not fall into the framework of Chapter 6 and b) this is achieves while being (mostly) adaptive to the loss range, rather than requiring an a-priori bound.[3]

Bandit Reduction and Variance Control To lift Theorem 40 to the contextual bandit setting with the ramp loss we use the following reduction: First, initialize the full information algorithm whose existence is guaranteed by Theorem 40 with $\mathcal{G} = \phi^\gamma \circ \mathcal{F}$. For each round t, receive x_t, and define $P_t(a) = \mathbb{E}_{s_t \sim p_t} \frac{s_t(a)}{\sum_{a' \in [K]} s_t(a')}$ where p_t is the full information algorithm's distribution. Then sample $a_t \sim P_t^\mu$, observe $\ell_t(a_t)$, and pass the importance-weighted loss

[2] Measuring loss in ℓ_1 may seem restrictive, but it is natural when working with the 1-sparse importance-weighted losses, and it enables us to cover the output space in ℓ_∞ norm.

[3] After optimizing the parameters λ and η in Theorem 40, the parameter R only enters a single term.

$\hat{\ell}_t(a)$ to the full information algorithm. For the hinge loss we use the same strategy, but with $\mathcal{G} = \psi^\gamma \circ \mathcal{F}$.

The following lemma shows that this strategy leads to sufficiently small variance in the loss estimates. The definition of the action distribution $P_t^\mu(a)$ in terms of the real-valued predictions is crucial here.

Lemma 32. Define a filtration $\mathcal{J}_t = \sigma((x_1, \ell_1, a_1), \ldots, (x_{t-1}, \ell_{t-1}, a_{t-1}), x_t, \ell_t)$. Then for any $\mu \in [0, 1/K]$ the importance weighting strategy above guarantees

$$\mathop{\mathbb{E}}_{a_t \sim P_t^\mu}\left[\mathop{\mathbb{E}}_{s_t \sim p_t}\left\langle s_t, \hat{\ell}_t\right\rangle^2 \mid \mathcal{J}_t\right] \leq \begin{cases} K, & \text{for } \mathcal{S} \subset \Delta(\mathcal{A}). \\ K^2, & \text{for } \mathcal{S} = \phi^\gamma \circ \mathcal{F}. \\ \left(1 + \frac{B}{\gamma}\right)^2 K^2, & \text{for } \mathcal{S} = \psi^\gamma \circ \mathcal{F}. \end{cases}$$

Theorem 44 and Lemma 32 together imply our central theorem: a chaining-based margin bound for contextual bandits, generalizing classical results in statistical learning (cf. (Boucheron et al., 2005)).

Theorem 41 (Contextual bandit margin bound)**.** *For any fixed constants $\beta > \alpha > 0$, smoothing parameter $\mu \in (0, 1)$ and margin loss parameter $\gamma > 0$ there exists an adversarial contextual bandit strategy with expected regret against the γ-margin benchmark bounded as*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^n \ell_t(a_t)\right] \leq & \inf_{f \in \mathcal{F}} \mathbb{E}[L_n^\gamma(f)] + 4\sqrt{2K^2 n \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{F}, n)} + \mu K n \\ & + \frac{8}{\mu} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{F}, n) + \frac{1}{\gamma}\left(3e^2 \alpha K n + 12e\sqrt{\frac{Kn}{\mu}} \int_\alpha^\beta \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n)} d\varepsilon\right).\end{aligned} \tag{11.3}$$

We derive an analogous bound based on the hinge loss, but since this is strictly weaker we defer the result to Section 11.2.

Before showing the implications of Theorem 41 for specific classes $\mathcal{F}$ we state a coarse upper bound in terms of the growth rate for the sequential metric entropy.

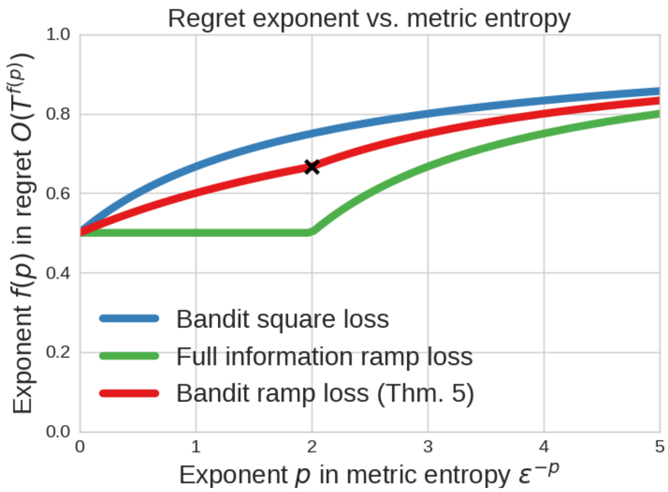


Figure 11.1: Regret bound exponent as a function of (sequential) metric entropy. The cross marks the point $p = 2$ where the exponent from Theorem 41 changes growth rate. "Full information" refers to the optimal rate of $n^{\frac{1}{2} \vee (\frac{p-1}{p})}$ for the same setting under full information feedback (Rakhlin et al., 2014). "Square loss" refers to the optimal rate of $n^{\frac{p+1}{p+2}}$ for Lipschitz contextual bandits over metric spaces of dimension p, which have sequential metric entropy ε^{-p}, under square loss realizability (Slivkins, 2011).

Proposition 19. Suppose that $\mathcal{F}$ has sequential metric entropy growth $\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n) \propto \varepsilon^{-p}$ for some $p > 0$ (nonparametric case), or that $\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n) \propto d \log(1/\varepsilon)$ (parametric case). Then there exists a contextual bandit strategy with the following regret guarantee:

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell_t(a_t)\right] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[L_n^{\gamma}(f)] + \begin{cases} O(K\sqrt{dn \log(Kn/\gamma)}), & \text{parametric case.} \\ \tilde{O}((Kn)^{\frac{p+2}{p+4}} \gamma^{-\frac{2p}{p+4}}), & \text{nonparametric w/ } p \leq 2. \\ \tilde{O}((Kn)^{\frac{p}{p+1}} \gamma^{-\frac{p}{p+1}}), & \text{nonparametric w/ } p \geq 2. \end{cases} \tag{11.4}$$

Proposition 19 recovers the parametric rate of $\sqrt{dn}$ seen with e.g., LinUCB (Chu et al., 2011) but is most interesting for complex classes. The rate exhibits a phase change between the "moderate complexity" regime of $p \in (0, 2]$ and the "high complexity" regime of $p \geq 2$. This is visualized in Figure 11.1.

Remark 4. *Under i.i.d. losses and hinge/ramp loss realizability, the standard tools of classification calibration (Bartlett et al., 2006) can be used to deduce a proper policy regret bound from (11.3). However, these realizability assumptions are somewhat non-standard, and moreover if one imposes the stronger assumption of a hard margin it is possible to derive improved rates (Daniely and Helbertal, 2013).*

Remark 5. *Compared to classical margin bounds which typically hold for all values of γ simultaneously, Theorem 41 requires that γ is chosen in advance. Learning the best value of γ online appears challenging.*

We now instantiate our results for concrete classes of interest.

Example 24 (Finite classes)**.** *In the finite class case there is an algorithm with $O\left(K\sqrt{n\log|\mathcal{F}|}\right)$ margin regret. When $\Pi \subset (\mathcal{X} \to \mathcal{A})$ is a finite policy class, our reduction to Theorem 40 yields the optimal $O\left(\sqrt{Kn\log|\Pi|}\right)$ policy regret, hinting at the optimality of our approach.*

Example 25 (Lipschitz CB)**.** *The class of all bounded Lipschitz functions over $[0,1]^p$ admits a pointwise cover with metric entropy $\tilde{O}(\varepsilon^{-p})$, immediately yielding a sequential cover. Proposition 19 thus implies an $\tilde{O}(n^{\frac{p+2}{p+4}\vee\frac{p}{p+1}})$ regret bound. Since our proof goes through Lemma 31, it also yields a* policy regret *bound against the $\pi_{ramp}(\cdot)$ policy class. Therefore, the result is directly comparable to the $\tilde{O}(n^{\frac{p+1}{p+2}})$ regret bound of Cesa-Bianchi et al. (2017) for Lipschitz contextual bandits (applied to the induced π_{ramp} policy class). Our bound achieves a smaller exponent for all values of p (see Figure 11.1).*

Learnability in the full information online learning setting is known to be characterized entirely by the sequential Rademacher complexity of the hypothesis class, and tight bounds on the sequential Rademacher complexity are known for standard classes including linear predictors, decision trees, and neural networks (Rakhlin et al., 2014). The next example, a corollary of Theorem 41, bounds contextual bandit margin regret in terms of sequential Rademacher complexity.

For any scalar-valued function class $\mathcal{G} \subseteq (\mathcal{X} \to \mathbb{R})$, define the sequential Rademacher complexity via

$$\mathcal{R}^{\mathrm{seq}}(\mathcal{G}) = \sup_{\boldsymbol{x}} \mathbb{E}_{\epsilon} \sup_{g\in\mathcal{G}} \sum_{t=1}^{n} \epsilon_t g(\boldsymbol{x}_t(\epsilon)).$$

Example 26. *Let* $\mathcal{F}|_a := \{x \mapsto f(x)_a \mid f \in \mathcal{F}\}$ *be the scalar restriction of* $\mathcal{F}$ *to output coordinate* a *and suppose that* $\max_{a \in [K]} \mathcal{R}^{\text{seq}}(\mathcal{F}|_a) \geq 1$ *and* $B \leq 1$.[4] *Then there exists an adversarial contextual bandit algorithm with margin regret bound* $\tilde{O}\Big(\max_a K(\mathcal{R}(\mathcal{F}|_a)/\gamma)^{2/3} n^{1/3}\Big)$.

This example implies that for margin-based contextual bandits, full information learnability is equivalent to bandit learnability. In particular, since the optimal regret in full information is $\Omega(\max_a \mathcal{R}^{\text{seq}}(\mathcal{F}|_a,))$, it further shows that the price of bandit information is at most $\tilde{O}\Big(\max_a K(n/\mathcal{R}^{\text{seq}}(\mathcal{F}|_a))^{1/3}\Big)$. Note however that while this bound is fairly user-friendly, the rates it obtains by plugging in the sequential metric entropy upper bound on sequential Rademacher complexity (Rakhlin et al., 2010) are sub-optimal relative to Proposition 19 except when $p = 2$. As a point of comparison, BISTRO (Rakhlin and Sridharan, 2016a) has an $O(\sqrt{Kn\mathcal{R}^{\text{seq}}(\Pi)})$ regret bound, which involves the complexity of the *policy* class (rather than the regressor class) and a worse n dependence than our bound, but our bound (in terms of $\mathcal{F}$) applies only to the margin regret.

We now instantiate Example 26 with linear classes. The next example generalizes the $O(n^{2/3})$ dimension-independent surrogate regret bound of the BANDITRON algorithm (Kakade et al., 2008) from Euclidean geometry to arbitrary uniformly convex Banach spaces (and more generally to Banach spaces of cotype 2), essentially the largest class of linear predictors for which online learning is possible (Srebro et al., 2011). The result also generalizes BANDITRON from bandit multiclass to general contextual bandits, and strengthens it from hinge loss to ramp loss. Note that many subsequent works (Abernethy and Rakhlin, 2009; Beygelzimer et al., 2017; Foster et al., 2018b) have obtained dimension-dependent $O(\sqrt{dn})$ bounds for bandit multiclass prediction, as we will in the next section, but, to our knowledge, none have explored dimension-independent $O(n^{2/3})$-type rates, which are more appropriate for high-dimensional settings.

[4] This restriction serves only to simplify calculations and can be relaxed.

Example 27. *Take $\mathcal{X}$ to the unit ball in a Banach space $(\mathfrak{B}, \|\cdot\|)$, and let $\mathcal{F}$ be the class of regressors induced by stacking $K-1$*[5] *linear predictors each in the unit ball of the dual Banach space $(\mathfrak{B}^{\star}, \|\cdot\|_{\star})$. Suppose that $\|\cdot\|$ has martingale type 2 (Pisier, 1975), which means there exists $\Psi : \mathfrak{B} \to \mathbb{R}$ such that $\frac{1}{2}\|x\|^2 \leq \Psi(x)$ and Ψ is β-smooth with respect to $\|\cdot\|$. Then there exists a contextual bandit strategy with margin regret $O(K(n/\gamma)^{2/3})$. Norms that satisfy the smoothness property with dimension-independent or logarithmic constants include ℓ_p for all $p \geq 2$, Schatten S_p norms for $p \geq 2$ (including the spectral norm), and $(2, p)$ group norms for $p \geq 2$ (Kakade et al., 2009b, 2012).*

Appealing to existing sequential Rademacher complexity bounds, we derive regret bounds for a few more well-known function classes. In the interest of space, we state these bounds informally and refer the reader to Rakhlin et al. (2014) for quantitative bounds on the sequential Rademacher complexity.

Example 28. *Suppose each $\mathcal{F}|_a$ consists of a class of neural networks with weights in each layer bounded in the $(1, \infty)$ group norm, or consists of a class of bounded depth decision trees with a finite set of decision functions. Then there exists a strategy with margin regret $\tilde{O}(K(n/\gamma)^{2/3})$.*

As our last example, we consider ℓ_p spaces for $p < 2$. These spaces fail to satisfy martingale type 2 in a dimension-independent fashion, but they do satisfy martingale type p without dimension dependence, and so have sequential metric entropy of order $\varepsilon^{-\frac{p}{p-1}}$ (Rakhlin and Sridharan, 2017). On the other hand, in $\mathbb{R}^d$ the ℓ_p spaces also admit a pointwise cover with metric entropy $O(d \log(1/\varepsilon))$, leading to the following dichotomy.

Example 29. *Consider the setting of Example 27, with $(\mathfrak{B}, \|\cdot\|) = (\mathbb{R}^d, \|\cdot\|_p)$ for $p \leq 2$. Then there exists a contextual bandit strategy with margin regret $\tilde{O}(K(n/\gamma)^{\frac{p}{2p-1}} \wedge K\sqrt{dn \log(Kn/\gamma)})$.*

[5]Only $K-1$ predictors are needed due to the sum-to-zero constraint of $\mathbb{R}^K_{=0}$.

11.3 Efficient Algorithms

This section contains two new algorithms for contextual bandits, both using the hinge loss ψ^γ. The first algorithm, HINGE-LMC, focuses on the parametric setting; it is based on a continuous version of exponential weights using a log-concave sampler, and is described in Section 11.3.1. The second, SMOOTHFTL, is simply Follow-The-Leader with uniform smoothing, described in Section 11.3.2. SMOOTHFTL applies to the stochastic contextual bandit setting with classes that have "high complexity" in the sense of Proposition 19.

Compared to the achievability results in this chapter, the algorithms presented in this section do not immediately arise from the equivalence framework of Part II. Whether the Burkholder method can be extended to generically solve contextual bandits and related problems with partial information remains an important open question.

11.3.1 Hinge-LMC

For this section, we identify $\mathcal{F}$ with a compact convex set $\Theta \subset \mathbb{R}^d$, using the notation $f(x;\theta) \in \mathbb{R}^K_{=0}$ to describe the parametrized function. We assume that $\psi^\gamma(f(x;\theta)_a)$ is convex in θ for each (x,a) pair, $\sup_{x,\theta} \|f(x;\theta)\|_\infty \leq B$, $f(x;\cdot)_a$ is L-Lipschitz as a function of θ with respect to the ℓ_2 norm, and that Θ contains the centered Euclidean ball of radius 1 and is contained within a Euclidean ball of radius R. These assumptions are all satisfied when $\mathcal{F}$ is a class of linear functions, under appropriate boundedness.

The pseudocode for the algorithm, HINGE-LMC, is displayed in Algorithm 12, with a sampling subroutine in Algorithm 13. The settings for all parameters are given in Section 11.6.1. The main idea is to run a continuous variant of exponential weights (Auer et al., 2002b) on the surrogate hinge loss. At round t, we define the exponential weights

Algorithm 12 HINGE-LMC

Input: Class Θ, learning rate η, rounds T, margin parameter γ.
Define $w_0(\theta) = 1$ for all $\theta \in \Theta$.
for $t = 1, \ldots, n$ **do**
  Receive x_t
  `// See Section 11.6.1 for LMC params.`
  $\theta_t \leftarrow \mathrm{LMC}(\eta w_{t-1})$.
  Set $p_t(\cdot; \theta_t) \propto \psi^\gamma(f(x_t; \theta_t))$
  Set $p_t^\mu(\cdot; \theta_t) = (1 - K\mu)p_t + \mu$.
  Play $a_t \sim p_t^\mu(\cdot; \theta_t)$, observe $\ell_t(a_t)$.
  `// Geometric resampling.`
  for $m = 1, \ldots, M$ **do**
    $\tilde{\theta}_t \leftarrow \mathrm{LMC}(\eta w_{t-1})$.
    Sample $\tilde{a}_t \sim p_t^\mu(\cdot; \tilde{\theta}_t)$, if $\tilde{a}_t = a_t$, break
  end for
  Set $m_t = m$, and $\tilde{\ell}_t(a) = \ell_t(a_t) \cdot m_t \mathbf{1}\{a_t = a\}$
  Update $w_t(\theta) \leftarrow w_{t-1}(\theta) + \langle \tilde{\ell}_t, \psi^\gamma(f(x_t; \theta)) \rangle$
end for

Algorithm 13 Langevin Monte Carlo (LMC)

Input: Function F, parameters m, u, λ, N, α.
Set $\tilde{\theta}_0 \leftarrow 0 \in \mathbb{R}^d$
for $k = 1, \ldots, N$ **do**
  Draw $z_1, \ldots, z_m \overset{iid}{\sim} \mathcal{N}(0, u^2 I_d)$ and define

$$\tilde{F}_k(\theta) = \frac{1}{m} \sum_{i=1}^m F(\theta + z_i) + \frac{\lambda}{2} \|\theta\|_2^2$$

  Draw $\xi_k \sim \mathcal{N}(0, I_d)$ and update

$$\tilde{\theta}_k \leftarrow \mathcal{P}_\Theta \left(\tilde{\theta}_{k-1} - \frac{\alpha}{2} \nabla \tilde{F}_k(\tilde{\theta}_{k-1}) + \sqrt{\alpha} \xi_k \right).$$

end for
Return $\tilde{\theta}_N$.

distribution via its density (w.r.t. the Lebesgue measure over Θ)

$$P_t(\theta) \propto \exp(-\eta w_{t-1}(\theta)), \qquad w_{t-1}(\theta) = \sum_{s=1}^{t-1} \langle \tilde{\ell}_s, \psi^\gamma(f(x_s; \theta)) \rangle,$$

where η is a learning rate and $\tilde{\ell}_s$ is an estimate of the loss vector. At a high level, at each iteration the algorithm generates a sample $\theta_t \sim P_t$, then samples the action a_t from the induced policy distribution $p_t(\cdot; \theta) = \pi_{\text{hinge}}(f(x_t; \theta_t)) \propto \psi^\gamma(f(x_t; \theta_t)) \in \Delta(\mathcal{A})$. The algorithm then plays a_t, observes the loss $\ell_t(a_t)$, and constructs a loss vector estimate $\tilde{\ell}_t = m_t \cdot \ell_t(a) \mathbf{1}\{a = a_t\}$, where m_t is an approximation to the importance weight computed by repeatedly sampling from P_t. This vector $\tilde{\ell}_t$ is passed to the exponential weights subroutine to define the distribution at the next round. To generate samples $\theta_t \sim P_t$ we use Projected Langevin Monte Carlo (LMC).

The algorithm has many important subtleties. Briefly, the analysis for Projected LMC that we use, due to Bubeck et al. (2018), requires a smooth potential function, and we use the randomized technique of Duchi et al. (2012) to smooth the hinge loss by convolving with a gaussian density (in expectation). Then, since the gradients of this smooth function cannot be computed in closed form, we use a coupling argument to show that the iterates on a sampled approximation track the ideal iterates. Here, the ℓ_2 regularization added to the function $\tilde{F}_k(\theta)$ defined in Algorithm 13 plays an important role. Finally, since we lack direct access to the sampling distribution, we use the geometric resampling technique of Neu and Bartók (2013) to approximate the importance weight by repeated sampling. At all stages it is important to show that the large scaling on the loss estimates induced by importance weighting does not degrade computational performance. All of the components are analyzed in detail in Section 11.6.1.

Here, we state the main guarantee and its consequences. A more complete theorem statement, with exact parameter specifications and the precise running time is provided in Section 11.6.1 as Theorem 46.

Theorem 42 (Informal)**.** *Under the assumptions of Section 11.3.1,* HINGE-LMC *with appropriate parameter settings runs in time poly*$(n, d, B, K, \frac{1}{\gamma}, R, L)$ *and guarantees*

$$\mathbb{E}\sum_{t=1}^{n}\ell_t(a_t) \leq \inf_{\theta\in\Theta}\frac{1}{K}\mathbb{E}\sum_{t=1}^{n}\langle \ell_t, \psi^{\gamma}(f(x_t;\theta))\rangle + \tilde{O}\left(\frac{B}{\gamma}\sqrt{dn}\right).$$

Since bandit multiclass prediction is a special case of contextual bandits, Theorem 42 immediately implies a $\sqrt{dn}$-mistake bound for this setting.

Corollary 13 (Bandit multiclass)**.** In the bandit multiclass setting, Algorithm 12 enjoys a mistake bound of $\tilde{O}((B/\gamma)\sqrt{dn})$ against the multiclass γ-hinge loss and runs in polynomial time.

Additionally, under a realizability condition for the hinge loss, we obtain a standard regret bound. For simplicity in defining the condition, assume that for every (x, ℓ) pair, ℓ is a random variable with conditional mean $\bar{\ell}$ (chosen by the adversary) and $\bar{\ell}$ has a unique action with minimal loss.

Corollary 14 (Realizable bound)**.** In addition to the conditions above, assume that there exists $\theta^\star \in \Theta$ such that for every (x, ℓ) pair and for all $a \in \mathcal{A}$, we have $f(x;\theta^\star)_a = K\gamma \mathbf{1}\{\bar{\ell}(a) \leq \min_{a'} \bar{\ell}(a')\} - \gamma$. Then HINGE-LMC runs in polynomial time and guarantees

$$\sum_{t=1}^{n} \mathbb{E}\bar{\ell}_t(a_t) \leq \sum_{t=1}^{n} \mathbb{E}\min_a \bar{\ell}(a) + \tilde{O}\left(\frac{B}{\gamma}\sqrt{dn}\right).$$

A few comments are in order:

1. On a technical level, apart from passing to the hinge surrogate loss to obtain a tractable log-concave sampling problem, the key insight is that the hinge loss also lets us controls the local norm term in the exponential weights regret bound (the first term on the right hand side of (11.1)). For this step, it is crucial that we sample from the induced policy distributions $\pi_{\text{hinge}}(\cdot)$ rather than the more natural argmax policy $\arg\max_a f(x;\theta)_a$, which does not provide suitable control. Our technique therefore seems specialized to surrogates that can be expressed as an inner product between the loss vector and (a transformation of) the prediction, which cannot be done for many loss functions used in bandit multiclass prediction.
2. The use of LMC for sampling is not strictly necessary. Other log-concave samplers do exist for non-smooth potentials (Lovász and Vempala, 2007), which will remove the parameters m, u, λ, significantly simplify the algorithm, and even lead to a better run-time guarantee using current theory. On the other hand, we prefer to use LMC due to its success in Bayesian inference and deep learning, and its connections to incremental optimization methods for supervised learning. LMC, moreso than say

Hit-and-Run (Lovász and Vempala, 2007), can easily be adapted to work quickly (in practice) when data arrives online. Furthermore, while the runtime for LMC is quite large in theory (Section 11.6.1), the theoretical memory usage scales only linearly with the memory required to store a single context. We are hopeful that the LMC approach will lead to a practically useful contextual bandit algorithm and plan to explore this direction further.

3. As mentioned above, while the algorithm is guaranteed to run in polynomial time, the dependence on n and d that we obtain is quite poor. In part, this inefficiency stems from the mixing-time analysis for Projected LMC (Bubeck et al., 2018). More recent results in slightly different settings (Raginsky et al., 2017; Dalalyan and Karagulyan, 2017; Cheng et al., 2018) suggest that it may be possible to substantially improve this analysis and even extend to non-convex settings. Similarly, we conjecture that Projected LMC can be analyzed without smoothness.
4. Corollary 13 provides a new solution to the open problem of Abernethy and Rakhlin (2009). In fact, this result is the first efficient $\sqrt{dn}$-type regret bound against a hinge loss benchmark, although it is slightly different from the multiclass hinge loss variant used by Kakade et al. (2008) in their $n^{2/3}$-regret BANDITRON algorithm (which was the motivation behind the open problem). All prior $\sqrt{dn}$-regret algorithms (Hazan and Kale, 2011; Beygelzimer et al., 2017; Foster et al., 2018b) use losses with curvature such as the multiclass logistic loss or the squared hinge loss.
5. In Corollary 14, regret is measured relative to the policy that chooses the best action (in expectation) on *every round*. As in prior results (Abbasi-Yadkori et al., 2011; Agarwal et al., 2012), this is possible because the realizability condition ensures that this policy is in our class. Note that here, a requirement for realizability is that $B \geq K\gamma$, and hence the dependence on K is implicit and in fact slightly worse than the optimal rate (Chu et al., 2011).

6. For Corollary 14, the best points of comparisons are methods based on square-loss realizability (Agarwal et al., 2012; Foster et al., 2018a), although our condition is different. We impose stronger assumptions on the regressor class but obtain better regret guarantees than those in Foster et al. (2018a), which is the only other efficient approach at a comparable level of generality. Our assumptions are somewhat weaker than for LINUCB and variants (Chu et al., 2011; Abbasi-Yadkori et al., 2011) that are specialized to the ℓ_2/ℓ_2 geometry,[6] but these methods have slightly better guarantees for linear classes (again under square loss realizability).

To summarize, HINGE-LMC is the first efficient $\sqrt{dn}$-regret algorithm for bandit multiclass prediction using the hinge loss. It also represents a new approach to adversarial contextual bandits, in which we obtain $\sqrt{dn}$ policy regret under hinge-based realizability. Finally, while we lose the theoretical guarantees, the algorithm easily extends to non-convex classes, which we expect to be practically effective.

11.3.2 SMOOTHFTL

A drawback of HINGE-LMC is that it only applies in the parametric regime. We now introduce an efficient (in terms of queries to a hinge loss minimization oracle) algorithm that enjoys a regret bound similar to Theorem 41, but under the additional assumption that data is stochastic. Precisely, we work in the same model as Section 11.1.1, but where the pairs $\{(x_t, \ell_t)\}_{t=1}^n$ are drawn i.i.d. from some joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathbb{R}_+^K$. For simplicity, we assume $B = 1$.

The algorithm we analyze is simply Follow-The-Leader with uniform smoothing and

[6] In abstract linear setting we take $\mathcal{F}$ to be the set of linear functions in the ball for some norm $\|\cdot\|$ and contexts to be bounded in the dual norm $\|\cdot\|_\star$. The runtime of HINGE-LMC will degrade (polynomially) with the ratio $\|\theta\|/\|\theta\|_2$, but the regret bound is the same for any such norm pair.

epoching, which we refer to as SMOOTHFTL. Here we return to the abstract setting with regression class $\mathcal{F}$. We use an epoch schedule where the m^{th} epoch lasts for $n_m = 2^m$ rounds (starting with $m = 0$). At the beginning of the m^{th} epoch, we compute the empirical importance weighted hinge-loss minimizer $\hat{f}_{m-1}$ using *only* the data from the previous epoch. That is, we set

$$\hat{f}_{m-1} = \arg\min_{f \in \mathcal{F}} \sum_{\tau=n_{m-1}}^{2n_{m-1}-1} \langle \hat{\ell}_\tau, \psi^\gamma(f(x_\tau)) \rangle.$$

Then, for each round t in the m^{th} epoch, we sample a_t according to $p_t = (1 - K\mu)\pi_{\text{hinge}}(\hat{f}_{m-1}(x_t)) + \mu$.

The parameter $\mu \in (0, 1/K]$ controls smoothing. At the first time $t = 1$ we simply take p_1 to be uniform.

Theorem 43 (SMOOTHFTL regret bound)**.** *Suppose that $\mathcal{F}$ satisfies $\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n) \propto \varepsilon^{-p}$ for some $p > 2$. Then in the stochastic setting, with $\mu = K^{-1} n^{\frac{-1}{p+1}}$,* SMOOTHFTL *enjoys the following expected regret guarantee*[7]

$$\sum_{t=1}^{n} \mathbb{E}\ell_t(a_t) \leq \inf_{f \in \mathcal{F}} \frac{n}{K} \mathbb{E}\langle \ell, \psi^\gamma(f(x)) \rangle + \tilde{O}\Big((n/\gamma)^{\frac{p}{p+1}} \Big).$$

This provides an algorithmic counterpart to Proposition 19 in the $p \geq 2$ regime. The algorithm is quite similar to EPOCH-GREEDY (Langford and Zhang, 2008), and the main contribution here is to provide a careful analysis for large function classes. We leave obtaining an oracle-efficient algorithm that matches Proposition 19 in the regime $p \in (0, 2)$ as an open problem.

Note that a similar bound can be obtained for the ramp loss by simply replacing the hinge loss ERM with that for the ramp loss. We analyze the hinge loss version because standard

[7] This result is stated in terms of the sequential cover $\mathcal{N}_{\infty,\infty}$ to avoid additional definitions, but can easily be improved to depend classical (worst-case) covering number seen in statistical learning.

(e.g. linear) classes admit efficient hinge loss minimization oracles. Interestingly, the bound in Theorem 43 actually improves on Proposition 19, in that it is independent of K. This is due to the scaling of the hinge loss in Lemma 31.

In Section 11.6.7, we extend the analysis to the stochastic Lipschitz contextual bandit setting. Here, instead of measuring regret against the benchmark $\psi^\gamma \circ \mathcal{F}$ we compare to the class of all 1-Lipschitz functions from $\mathcal{X}$ to $\Delta(\mathcal{A})$, where $\mathcal{X}$ is some metric space of bounded covering dimension. We show that SMOOTHFTL achieves $n^{\frac{p}{p+1}}$ regret against Lipschitz policies over a p-dimensional context space with finite action space. This improves on the $n^{\frac{p+1}{p+2}}$ bound of Cesa-Bianchi et al. (2017), as in Example 25, yet the best available lower bound is $n^{\frac{p-1}{p}}$ (Hazan and Megiddo, 2007). Closing this gap remains an intriguing open problem.

11.4 Discussion

This chapter initiates a study of the utility of surrogate losses in contextual bandit learning. We obtain new margin-based regret bounds in terms of sequential complexity notions on the benchmark class, improving on the best known rates for Lipschitz contextual bandits and providing dimension-independent bounds for linear classes. On the algorithmic side, we provide the first solution to the open problem of Abernethy and Rakhlin (2009) with a loss without curvature, and we also show that Follow-the-Leader with uniform smoothing performs well in nonparametric settings.

Yet, several open problems remain. First, our bounds are likely suboptimal in the dependence on K. Next, while Proposition 19 recovers all upper bounds we are aware of (e.g., the $O(T^{2/3})$ dimension-independent bound of BANDITRON), a matching lower bound is not available, and resolving the optimality of Proposition 19 in all regimes is an intriguing open problem. Another important problem is to adapt to the margin parameter—this is easy in the classical statistical learning setting via penalized risk minimization, but partial information makes adapting such an approach nontrivial.

11.5 Detailed Proofs for Minimax Results

11.5.1 Calibration Lemmas

Proof of Lemma 31. We start with the ramp loss. First since $s \in \mathbb{R}^K_{=0}$, we know that the normalization term in $\pi_{\text{ramp}}(s)$ is

$$\sum_{a \in \mathcal{A}} \phi^\gamma(s_a) \geq 1,$$

from which the first inequality follows. The second inequality follows from the fact that $s_a \leq -\gamma$ implies that $\pi_{\text{ramp}}(s)_a = 0$, along with the trivial fact that $\pi_{\text{ramp}}(s)_a \leq 1$.

The hinge loss claim is also straightforward, since here the normalization is

$$\sum_{a \in \mathcal{A}} \psi^\gamma(s_a) = \sum_{a \in \mathcal{A}} \max\{1 + s_a/\gamma, 0\} \geq \sum_{a} 1 + \frac{s_a}{\gamma} \geq K. \qquad \square$$

Lemma 33 (Hinge loss realizability)**.** Let $\ell \in \mathbb{R}^K_+$ and let $a^\star = \arg\min_{a \in \mathcal{A}} \ell_a$. Define $s \in \mathbb{R}^K_{=0}$ via $s_a \triangleq K\gamma \mathbf{1}\{a = a^\star\} - \gamma$. Then we have

$$\langle \ell, \psi^\gamma(s) \rangle = K \langle \ell, \pi_{\text{hinge}}(s) \rangle = K \ell_{a^\star}.$$

Proof. For this particular s, the normalizing constant in the definition of π_{hinge} is

$$\sum_{a \in \mathcal{A}} \max\left(1 + \frac{K\gamma \mathbf{1}\{a = a^\star\} - \gamma}{\gamma}, 0\right) = K,$$

and so the first equality follows. The second equality is also straightforward since the score for every action except $a^\star$ is clamped to zero. $\square$

Proof of Lemma 32.

For the case when $\mathcal{S} \subset \Delta(\mathcal{A})$, this claim is a well-known property of importance weighting:

$$\mathbb{E}\left[\mathop{\mathbb{E}}_{s_t\sim p_t}\left\langle s_t, \hat{\ell}_t\right\rangle^2 \mid \mathcal{J}_t\right] = \sum_{a\in[K]} P_t^\mu(a)\frac{\mathbb{E}_{s_t\sim p_t}\,\ell_t^2(a)s_t^2(a)}{(P_t^\mu(a))^2} \leq \sum_{a\in[K]} \frac{\mathbb{E}_{s_t\sim p_t}s_t^2(a)}{P_t^\mu(a)}$$
$$\leq \sum_{a\in[K]} \frac{\mathbb{E}_{s_t\sim p_t}s_t(a)}{P_t^\mu(a)} = \sum_{a\in[K]} \frac{P_t(a)}{(1-K\mu)P_t(a)+\mu}.$$

Here we use Hölder's inequality twice, using that $\|\ell\|_\infty \leq 1$ and $s \in \Delta(\mathcal{A})$. Now, since the function $x \mapsto 1/(1-K\mu+\mu/x)$ is concave in x, it follows that

$$\sum_{a\in[K]} \frac{P_t(a)}{(1-K\mu)P_t(a)+\mu} = \sum_{a\in[K]} \frac{1}{(1-K\mu)+\mu/P_t(a)}$$
$$\leq K\frac{1}{(1-K\mu)+K\mu/\sum_{a\in[K]}P_t(a)} = K,$$

which proves the claim for $\mathcal{S} \subset \Delta(\mathcal{A})$.

We proceed in the same fashion for both the ramp and hinge loss. Recall the definition $P_t^\mu(a) = (1-K\mu)\mathbb{E}_{s_t\sim p_t}\frac{s_t(a)}{\sum_{a'\in[K]}s_t(a')} + \mu$. We have

$$\mathbb{E}\left[\mathop{\mathbb{E}}_{s_t\sim p_t}\left\langle s_t, \hat{\ell}_t\right\rangle^2 \mid \mathcal{J}_t\right] = \sum_{a\in[K]} P_t^\mu(a)\frac{\mathbb{E}_{s_t\sim p_t}\,\ell_t^2(a)s_t^2(a)}{(P_t^\mu(a))^2} = \sum_{a\in[K]} \frac{\mathbb{E}_{s_t\sim p_t}\,\ell_t^2(a)s_t^2(a)}{P_t^\mu(a)}$$
$$\leq \sum_{a\in[K]} \frac{\mathbb{E}_{s_t\sim p_t}\,s_t^2(a)}{P_t^\mu(a)} \leq \max_{a\in[K]}\max_{s\in\mathcal{S}} s(a)\cdot\sum_{a\in[K]}\frac{\mathbb{E}_{s_t\sim p_t}\,s_t(a)}{P_t^\mu(a)}$$
$$= \max_{a\in[K]}\max_{s\in\mathcal{S}} s(a)\cdot\sum_{a\in[K]}\frac{\mathbb{E}_{s_t\sim p_t}\,s_t(a)}{(1-\mu K)\,\mathbb{E}_{s_t\sim p_t}\frac{s_t(a)}{\sum_{a'\in[K]}s_t(a')}+\mu}$$
$$\leq K\cdot\left(\max_{a\in[K]}\max_{s\in\mathcal{S}} s(a)\right)^2\cdot\sum_{a\in[K]}\frac{\mathbb{E}_{s_t\sim p_t}\frac{s_t(a)}{\sum_{a'\in[K]}s_t(a')}}{(1-K\mu)\,\mathbb{E}_{s_t\sim p_t}\frac{s_t(a)}{\sum_{a'\in[K]}s_t(a')}+\mu}.$$

Here we first apply the definition of $\hat{\ell}_t$ and cancel out one factor of P_t^μ in the denomator. Then we apply Hölder's inequality, using that $s_t(a) \geq 0$. Expanding the definition P_t^μ and using the upper bound $\sum_{a'\in[K]}s_t(a') \leq K\max_a\max_s s_t(a)$, yields the final expression.

Now, let $q_a \triangleq \mathbb{E}_{s_t \sim p_t} \frac{s_t(a)}{\sum_{a' \in [K]} s_t(a')}$, and apply the concavity argument above. This yields

$$K^2 \cdot \left(\max_{a \in [K]} \max_{s \in \mathcal{S}} s(a) \right)^2.$$

For the set $\mathcal{S}$ induced by the ramp loss we have $\max_{a \in [K]} \max_{s \in \mathcal{S}} s(a) \leq 1$, and for the set $\mathcal{S}$ induced by the hinge loss we have $\max_{a \in [K]} \max_{s \in \mathcal{S}} s(a) \leq (1 + \frac{B}{\gamma})$. □

11.5.2 Proofs from Section 11.2

Let us start with an intermediate result, which will simplify the proof of Theorem 40.

Theorem 44. *Assume $\|\ell\|_1 \leq 1$ for all $\ell \in \mathcal{L}$*[8] *and $\sup_{s \in \mathcal{S}} \|s\|_\infty \leq 1$. Further assume that $\mathcal{S}$ and $\mathcal{L}$ are compact. Fix any constants $\eta \in (0, 1]$, $\lambda > 0$, and $\beta > \alpha > 0$. Then there exists an algorithm with the following deterministic regret guarantee:*

$$\begin{aligned}\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s, \ell_t \rangle - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle \leq 2\eta \sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s_t, \ell_t \rangle^2 + \frac{4}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) + 3e^2 \alpha \sum_{t=1}^{n} \|\ell_t\|_1 \\ + 12e \left(\frac{\lambda}{4} \sum_{t=1}^{n} \|\ell_t\|_1^2 + \frac{1}{\lambda} \right) \int_\alpha^\beta \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon.\end{aligned}$$

The difference here is that have set $R, B = 1$. The first part of this section will be devoted to proving this theorem, and Theorem 40 will follow from this result via Corollary 15.

11.5.3 Preliminaries

Definition 18 (Cover for a collection of trees)**.** *For a collection of $\mathbb{R}^K$-valued trees U of length n, we let $\mathcal{N}_{\infty,\infty}(\varepsilon, U)$, denote the cardinality of the smallest set V of $\mathbb{R}^K$ valued trees for which*

$$\forall \boldsymbol{u} \in U \ \forall \epsilon \in \{\pm 1\}^n \ \exists \boldsymbol{v} \in V \text{ s.t. } \max_{t \in [n]} \|\boldsymbol{u}_t(\epsilon) - \boldsymbol{v}_t(\epsilon)\|_\infty \leq \varepsilon.$$

[8] Measuring loss in ℓ_1 may seem restrictive, but this is natural when working with importance-weighted losses since these are 1-sparse, and by duality this enables us to cover in ℓ_∞ norm on the output space.

Definition 19 (L_∞/ℓ_∞ radius). *For a function class $\mathcal{F}$, define*

$$\mathrm{rad}_{\infty,\infty}(\mathcal{F}, n) = \min\{\varepsilon \mid \log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n) = 0\}.$$

For a collection U of trees, define $\mathrm{rad}_{\infty,\infty}(U) = \min\{\varepsilon \mid \log \mathcal{N}_{\infty,\infty}(\varepsilon, U) = 0\}$.

The following two lemmas are Freedman-type inequalities for Rademacher tree processes that we will use in the sequel. The first has an explicit dependence on the range, while the second does not.

Lemma 34. For any collection of $[-R, +R]$-valued trees V of length n, for any $\eta > 0$ and $\alpha > 0$,

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{v} \in V} \left[\sum_{t=1}^{n} \epsilon_t \big(\boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \big) - \alpha \eta \boldsymbol{v}_t^2(\epsilon) \right] \leq 2 \log|V| \cdot \left(\frac{1}{\alpha\eta} \vee \frac{\eta R^2}{\alpha} \right).$$

Proof of Lemma 34. Take V to be finite without loss of generality (otherwise the bound is vacuous). As a starting point, for any $\lambda > 0$ we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{v} \in V} \left[\sum_{t=1}^{n} \epsilon_t \big(\boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \big) - \alpha \eta \boldsymbol{v}_t^2(\epsilon) \right] \\
&\leq \frac{1}{\lambda} \log \left(\sum_{\boldsymbol{v} \in V} \mathop{\mathbb{E}}_{\epsilon} \exp \left(\sum_{t=1}^{n} \epsilon_t \lambda \big(\boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \big) - \lambda \alpha \eta \boldsymbol{v}_t^2(\epsilon) \right) \right).
\end{aligned}$$

Applying the standard Rademacher mgf bound $\mathbb{E}_{\epsilon} e^{\lambda\epsilon} \leq e^{\frac{1}{2}\lambda^2}$ conditionally at each time starting from $t = n$, this is upper bounded by

$$\leq \frac{1}{\lambda} \log \left(\sum_{\boldsymbol{v} \in V} \max_{\epsilon} \exp \left(\sum_{t=1}^{n} \frac{1}{2} \lambda^2 \big(\boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \big)^2 - \lambda \alpha \eta \boldsymbol{v}_t^2(\epsilon) \right) \right).$$

Since v takes values in $[-R, +R]$, the exponent at time t can be upper bounded as

$$\frac{1}{2} \lambda^2 \big(\boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \big)^2 - \lambda \alpha \eta \boldsymbol{v}_t^2(\epsilon) \leq \lambda^2 \big(1 + \eta^2 R^2\big) \boldsymbol{v}_t^2(\epsilon) - \lambda \alpha \eta \boldsymbol{v}_t^2(\epsilon).$$

By setting $\lambda = \frac{1}{2} \min\{\alpha\eta, \alpha/(\eta R^2)\}$, this is bounded by zero, which leads to a final bound of $\log|V|/\lambda$. $\square$

Lemma 35. For any collection of trees V of length n, for any $\eta > 0$,

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{v} \in V} \left[\sum_{t=1}^{n} \epsilon_t \boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \right] \leq \frac{\log|V|}{2\eta}.$$

Proof of Lemma 35. Take V to be finite without loss of generality. As in the proof of Lemma 34, using the standard Rademacher mgf bound and working backward from n, for any $\lambda > 0$ we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{v} \in V} \left[\sum_{t=1}^{n} \epsilon_t \boldsymbol{v}_t(\epsilon) - \eta \boldsymbol{v}_t^2(\epsilon) \right] &\leq \frac{1}{\lambda} \log \left(\sum_{\boldsymbol{v} \in V} \mathop{\mathbb{E}}_{\epsilon} \exp \left(\sum_{t=1}^{n} \epsilon_t \lambda \boldsymbol{v}_t(\epsilon) - \eta \lambda \boldsymbol{v}_t^2(\epsilon) \right) \right) \\
&\leq \frac{1}{\lambda} \log \left(\sum_{\boldsymbol{v} \in V} \max_{\epsilon} \exp \left(\sum_{t=1}^{n} \frac{1}{2} \lambda^2 \boldsymbol{v}_t(\epsilon)^2 - \eta \lambda \boldsymbol{v}_t^2(\epsilon) \right) \right).
\end{aligned}$$

The exponent at time t is

$$\frac{1}{2} \lambda^2 \boldsymbol{v}_t^2(\epsilon) - \eta \lambda \boldsymbol{v}_t^2(\epsilon).$$

By setting $\lambda = 2\eta$, this is exactly zero, which leads to a final bound of $\log|V|/\lambda$. $\square$

Lemma 36. Let $\mathcal{Z}$, $\mathcal{W}$, and $\mathcal{G}$ be abstract sets and let functions $A_g : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ and $B_g : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be given for each element $g \in \mathcal{G}$. Suppose that for any $z, z' \in \mathcal{Z}$ and $w \in \mathcal{W}$ it holds that $A(w, z, z') = -A(w, z', z)$ and $B(w, z, z') = B(w, z', z)$. Then

$$\left\langle\!\!\left\langle \sup_{w_t \in \mathcal{W}} \sup_{q_t \in \Delta(\mathcal{Z})} \mathop{\mathbb{E}}_{z_t, z_t' \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} A_g(w_t, z_t, z_t') + B_g(w_t, z_t, z_t') \tag{11.5}$$

$$\leq \left\langle\!\!\left\langle \sup_{w_t \in \mathcal{W}} \sup_{q_t \in \Delta(\mathcal{Z})} \mathop{\mathbb{E}}_{\epsilon_t} \mathop{\mathbb{E}}_{z_t, z_t' \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \epsilon_t A_g(w_t, z_t, z_t') + B_g(w_t, z_t, z_t'), \tag{11.6}$$

where ϵ is a sequence of independent Rademacher random variables.

Proof of Lemma 36. See proof of Lemma 3 in Rakhlin et al. (2010). $\square$

11.5.4 Proof of Theorem 44

Let $\eta_1, \eta_2, \eta_3 > 0$ be fixed constants to be chosen later in the proof, and define

$$B(p_{1:n}, \ell_{1:n}) \triangleq \underbrace{\eta_1 \sum_{t=1}^{n} \|\ell_t\|_1 + \eta_2 \sum_{t=1}^{n} \|\ell_t\|_1^2}_{\triangleq B_1(\ell_{1:n})} + \underbrace{2\eta_3 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{s\sim p_t} \langle s, \ell_t \rangle^2}_{\triangleq B_2(p_{1:n}, \ell_{1:n})}.$$

We consider a game where the goal of the learner is to achieve regret bounded by B, plus some additive constant that will depend on η_1, η_2, η_3, and the complexity of the class $\mathcal{F}$. The minimax achievability of B is given by

$$\mathcal{V}_n^{\mathrm{ol}}(\mathcal{G}, B) \triangleq \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{p_t \in \Delta(\mathcal{S})} \sup_{\ell_t \in \mathcal{L}} \mathop{\mathbb{E}}_{s \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \langle s, \ell_t \rangle - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B(p_{1:n}, \ell_{1:n}) \right].$$

Following discussion in Chapter 2, if we show that $\mathcal{V}_n^{\mathrm{ol}}(\mathcal{G}, B) \leq C$ for some constant C then we have established existence of a randomized strategy that achieves an adaptive regret bound of $B(\cdot) + C$. Going forward we adopt the abbreviation $\mathcal{V}_n^{\mathrm{ol}} := \mathcal{V}_n^{\mathrm{ol}}(\mathcal{G}, B)$.

Minimax swap

At time t the value to go is given by

$$\begin{aligned}
&\sup_{x_t \in \mathcal{X}} \inf_{p_t \in \Delta(\mathcal{S})} \sup_{\ell_t \in \mathcal{L}} \Bigg[\mathop{\mathbb{E}}_{s \sim p_t} \langle s, \ell_t \rangle - 2\eta_3 \mathop{\mathbb{E}}_{s \sim p_t} \langle s, \ell_t \rangle^2 - \eta_1 \|\ell_t\|_1 - \eta_2 \|\ell_t\|_1^2 \\
&\quad + \left\langle\!\!\left\langle \sup_{x_\tau \in \mathcal{X}} \inf_{p_\tau \in \Delta(\mathcal{S})} \sup_{\ell_\tau \in \mathcal{L}} \right\rangle\!\!\right\rangle_{\tau=t+1}^{n} \left[\sum_{\tau=t+1}^{n} \mathop{\mathbb{E}}_{s \sim p_\tau} \langle s, \ell_\tau \rangle - \inf_{g \in \mathcal{G}} \sum_{\tau=1}^{n} \langle g(x_\tau), \ell_\tau \rangle - B(p_{\tau+1:n}, \ell_{\tau+1:n}) \right] \Bigg].
\end{aligned}$$

Note that the benchmark's loss is only evaluated at the end, while we are incorporating the adaptive term into the instantaneous value. Convexifying the ℓ_t player by allowing them to select a randomized strategy q_t, this is equal to

$$\begin{aligned}
&\sup_{x_t \in \mathcal{X}} \inf_{p_t \in \Delta(\mathcal{S})} \sup_{q_t \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t \sim q_t} \Bigg[\mathop{\mathbb{E}}_{s \sim p_t} \langle s, \ell_t \rangle - 2\eta_3 \mathop{\mathbb{E}}_{s \sim p_t} \langle s, \ell_t \rangle^2 - \eta_1 \|\ell_t\|_1 - \eta_2 \|\ell_t\|_1^2 \\
&\quad + \left\langle\!\!\left\langle \sup_{x_\tau \in \mathcal{X}} \inf_{p_\tau \in \Delta(\mathcal{S})} \sup_{\ell_\tau \in \mathcal{L}} \right\rangle\!\!\right\rangle_{\tau=t+1}^{n} \left[\sum_{\tau=t+1}^{n} \mathop{\mathbb{E}}_{s \sim p_\tau} \langle s, \ell_\tau \rangle - \inf_{g \in \mathcal{G}} \sum_{\tau=1}^{n} \langle g(x_\tau), \ell_\tau \rangle - B(p_{\tau+1:n}, \ell_{\tau+1:n}) \right] \Bigg].
\end{aligned}$$

This quantity is convex in p_t and linear in q_t so, under the compactness assumption on $\mathcal{S}$ and $\mathcal{L}$, the minimax theorem (Section 2.6) implies that this is equal to

$$\sup_{x_t \in \mathcal{X}} \sup_{q_t \in \Delta(\mathcal{L})} \inf_{p_t \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell_t \sim q_t} \Bigg[\mathop{\mathbb{E}}_{s \sim p_t} \langle s, \ell_t \rangle - 2\eta_3 \mathop{\mathbb{E}}_{s \sim p_t} \langle s, \ell_t \rangle^2 - \eta_1 \|\ell_t\|_1 - \eta_2 \|\ell_t\|_1^2 \\ + \left\langle\!\!\left\langle \sup_{x_\tau \in \mathcal{X}} \inf_{p_\tau \in \Delta(\mathcal{S})} \sup_{\ell_\tau \in \mathcal{L}} \right\rangle\!\!\right\rangle_{\tau=t+1}^{n} \left[\sum_{\tau=t+1}^{n} \mathop{\mathbb{E}}_{s \sim p_\tau} \langle s, \ell_\tau \rangle - \inf_{g \in \mathcal{G}} \sum_{\tau=1}^{n} \langle g(x_\tau), \ell_\tau \rangle - B(p_{\tau+1:n}, \ell_{\tau+1:n}) \right] \Bigg].$$

Repeating this analysis at each timestep and expanding the terms from B_2, we arrive at the expression

$$\mathcal{V}_n^{\mathrm{ol}} = \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{q_t \in \Delta(\mathcal{L})} \inf_{p_t \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s \sim p_t} \left[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \right] - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \right].$$

Upper bound by martingale process

We now use a standard "rearrangement" trick (see (Rakhlin et al., 2014), Theorem 1) to show that $\mathcal{V}_n^{\mathrm{ol}}$ is equal to

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{q_t \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \inf_{p_t \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{s \sim p_t} \mathop{\mathbb{E}}_{\ell'_t \sim q_t} \left[\langle s, \ell'_t \rangle - 2\eta_3 \langle s, \ell'_t \rangle^2 \right] - \sum_{t=1}^{n} \langle f(x_t), \ell_t \rangle \right] - B_1(\ell_{1:n}) \right],$$

where $\ell'_{1:n}$ is a sequence of "tangent" samples, where ℓ'_t is an independent copy of ℓ_t conditioned on $\ell_{1:t-1}$. This can be seen by working backwards from time n. Indeed, at time n, expanding the $\langle\!\langle \star \rangle\!\rangle_{t=1}^{n}$ operator, we have

$$\mathcal{V}_n^{\mathrm{ol}} = \langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-1} \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell_n \sim q_n} \Bigg[\sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{s \sim p_t} \left[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \right] + \mathop{\mathbb{E}}_{s \sim p_n} \left[\langle s, \ell_n \rangle - 2\eta_3 \langle s, \ell_n \rangle^2 \right] \\ - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg].$$

Using linearity of expectation this is equivalen to

$$\langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-1} \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell_n \sim q_n} \Bigg[\sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{s \sim p_t} \left[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \right] + \mathop{\mathbb{E}}_{\ell'_n \sim q_n} \mathop{\mathbb{E}}_{s \sim p_n} \left[\langle s, \ell'_n \rangle - 2\eta_3 \langle s, \ell'_n \rangle^2 \right] \\ - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg].$$

Using that only a single term has functional dependence on p_n, this is equal to

$$\langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-1} \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_n \sim q_n} \Bigg[\sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{s \sim p_t} \Big[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \Big] + \inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_n \sim q_n} \mathop{\mathbb{E}}_{s \sim p_n} \Big[\big\langle s, \ell'_n \big\rangle - 2\eta_3 \big\langle s, \ell'_n \big\rangle^2 \Big]$$
$$- \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg].$$

Expanding the infimum over $g \in \mathcal{G}$, this is equal to

$$\langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-1} \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_n \sim q_n} \sup_{g \in \mathcal{G}} \Bigg[\sum_{t=1}^{n-1} \mathop{\mathbb{E}}_{s \sim p_t} \Big[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \Big]$$
$$+ \inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_n \sim q_n} \mathop{\mathbb{E}}_{s \sim p_n} \Big[\big\langle s, \ell'_n \big\rangle - 2\eta_3 \big\langle s, \ell'_n \big\rangle^2 \Big] - \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg].$$

We handle time $n-1$ in a similar fashion by first splitting the $\langle\!\langle \star \rangle\!\rangle_{t=1}^{n-1}$ operator:

$$= \langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-2} \sup_{x_{n-1} \in \mathcal{X}} \sup_{q_{n-1} \in \Delta(\mathcal{L})} \inf_{p_{n-1} \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell_{n-1} \sim q_{n-1}} \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_n \sim q_n}$$
$$\sup_{g \in \mathcal{G}} \Bigg[\sum_{t=1}^{n-2} \mathop{\mathbb{E}}_{s \sim p_t} \Big[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \Big] + \mathop{\mathbb{E}}_{s \sim p_{n-1}} \Big[\langle s, \ell_{n-1} \rangle - 2\eta_3 \langle s, \ell_{n-1} \rangle^2 \Big]$$
$$+ \inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_n \sim q_n} \mathop{\mathbb{E}}_{s \sim p_n} \Big[\big\langle s, \ell'_n \big\rangle - 2\eta_3 \big\langle s, \ell'_n \big\rangle^2 \Big] - \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg].$$

Rearranging the supremums to make dependence on terms from time $n-1$ clear:

$$= \langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-2} \sup_{x_{n-1} \in \mathcal{X}} \sup_{q_{n-1} \in \Delta(\mathcal{L})} \inf_{p_{n-1} \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell_{n-1} \sim q_{n-1}}$$
$$\Bigg[\sum_{t=1}^{n-2} \mathop{\mathbb{E}}_{s \sim p_t} \Big[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \Big] + \mathop{\mathbb{E}}_{s \sim p_{n-1}} \Big[\langle s, \ell_{n-1} \rangle - 2\eta_3 \langle s, \ell_{n-1} \rangle^2 \Big]$$
$$+ \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_n \sim q_n} \sup_{g \in \mathcal{G}} \Bigg[\inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_n \sim q_n} \mathop{\mathbb{E}}_{s \sim p_n} \Big[\big\langle s, \ell'_n \big\rangle - 2\eta_3 \big\langle s, \ell'_n \big\rangle^2 \Big] - \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg] \Bigg].$$

Using linearity of expectation and moving the infimum over q_{n-1}:

$$= \langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-2} \sup_{x_{n-1} \in \mathcal{X}} \sup_{q_{n-1} \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_{n-1} \sim q_{n-1}}$$
$$\Bigg[\sum_{t=1}^{n-2} \mathop{\mathbb{E}}_{s \sim p_t} \Big[\langle s, \ell_t \rangle - 2\eta_3 \langle s, \ell_t \rangle^2 \Big] + \inf_{p_{n-1} \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_{n-1} \sim q_{n-1}} \mathop{\mathbb{E}}_{s \sim p_{n-1}} \Big[\big\langle s, \ell'_{n-1} \big\rangle - 2\eta_3 \big\langle s, \ell'_{n-1} \big\rangle^2 \Big]$$
$$+ \sup_{x_n \in \mathcal{X}} \sup_{q_n \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_n \sim q_n} \sup_{g \in \mathcal{G}} \Bigg[\inf_{p_n \in \Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_n \sim q_n} \mathop{\mathbb{E}}_{s \sim p_n} \Big[\big\langle s, \ell'_n \big\rangle - 2\eta_3 \big\langle s, \ell'_n \big\rangle^2 \Big] - \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle - B_1(\ell_{1:n}) \Bigg] \Bigg].$$

The last step is to move the supremums from time $t = n$ and the supremum over $g \in \mathcal{G}$ outside the entire expression, similar to what was done at time $t = n$.

$$
\begin{aligned}
&= \langle\!\langle \cdots \rangle\!\rangle_{t=1}^{n-2} \sup_{x_{n-1}\in\mathcal{X}} \sup_{q_{n-1}\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_{n-1}\sim q_{n-1}} \sup_{x_n\in\mathcal{X}} \sup_{q_n\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_n\sim q_n} \sup_{g\in\mathcal{G}} \\
&\qquad \left[\sum_{t=1}^{n-2} \mathop{\mathbb{E}}_{s\sim p_t}\left[\langle s, \ell_t\rangle - 2\eta_3 \langle s, \ell_t\rangle^2 \right] + \inf_{p_{n-1}\in\Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_{n-1}\sim q_{n-1}} \mathop{\mathbb{E}}_{s\sim p_{n-1}} \left[\left\langle s, \ell'_{n-1}\right\rangle - 2\eta_3 \left\langle s, \ell'_{n-1}\right\rangle^2 \right] \right. \\
&\qquad \left. + \inf_{p_n\in\Delta(\mathcal{S})} \mathop{\mathbb{E}}_{\ell'_n\sim q_n} \mathop{\mathbb{E}}_{s\sim p_n} \left[\left\langle s, \ell'_n\right\rangle - 2\eta_3 \left\langle s, \ell'_n\right\rangle^2 \right] - \sum_{t=1}^{n} \langle g(x_t), \ell_t\rangle - B_1(\ell_{1:n}) \right].
\end{aligned}
$$

Repeating this argument down from time $t = n - 2$ to time $t = 1$ yields the result.

To conclude this portion of the proof, we move to an upper bound by choosing the infimum over p_t at each timestep t to match g, which is possible because each infimum now occurs inside the expression for which the supremum over $g \in \mathcal{G}$ is taken. First, observe that the minimax value $\mathcal{V}_n^{\mathrm{ol}}$ is equal to

$$
\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}} \left[\sum_{t=1}^{n} \inf_{p_t\in\Delta(\mathcal{S})} \mathop{\mathbb{E}}_{s\sim p_t} \mathop{\mathbb{E}}_{\ell'_t\sim q_t} \left[\left\langle s, \ell'_t\right\rangle - 2\eta_3 \left\langle s, \ell'_t\right\rangle^2 \right] - \sum_{t=1}^{n} \langle f(x_t), \ell_t\rangle \right] - B_1(\ell_{1:n}) \right].
$$

Next, we have an upper bound of

$$
\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}} \left[\sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell'_t\sim q_t} \left[\left\langle g(x_t), \ell'_t\right\rangle - 2\eta_3 \left\langle g(x_t), \ell'_t\right\rangle^2 \right] - \sum_{t=1}^{n} \langle g(x_t), \ell_t\rangle \right] - B_1(\ell_{1:n}) \right].
$$

Finally, this is equal to

$$
\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}} \left[\sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell'_t\sim q_t} \left[\left\langle g(x_t), \ell'_t\right\rangle \right] - \langle g(x_t), \ell_t\rangle - 2\eta_3 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell'_t\sim q_t} \left[\left\langle g(x_t), \ell'_t\right\rangle^2 \right] \right] \right. \\
&\qquad \left. - B_1(\ell_{1:n}) \right].
\end{aligned}
\tag{11.7}
$$

Symmetrization

Introduce the notation $H(x) = x - \eta_3 x^2$. We now claim that the quantity appearing in (11.7) is bounded by

$$2 \cdot \sup_{\boldsymbol{x}} \sup_{\boldsymbol{\ell}} \mathbb{E}_{\epsilon}\left[\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} \epsilon_t H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon)\rangle) - \eta_3 \sum_{t=1}^{n}\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon)\rangle^2\right] - B_1(\boldsymbol{\ell}_{1:n}(\epsilon))\right], \tag{11.8}$$

where the supremum ranges over all $\mathcal{X}$-valued trees $\boldsymbol{x}$ and $\mathcal{L}$-valued trees $\boldsymbol{\ell}$, both of length n.

The value

$$\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathbb{E}_{\ell_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} \mathbb{E}_{\ell'_t\sim q_t}\left[\langle g(x_t), \ell'_t\rangle\right] - \langle g(x_t), \ell_t\rangle - 2\eta_3 \sum_{t=1}^{n} \mathbb{E}_{\ell'_t\sim q_t}\left[\langle g(x_t), \ell'_t\rangle^2\right]\right] - B_1(\ell_{1:n})\right],$$

by adding and subtracting the same term, is equal to

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathbb{E}_{\ell_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} \mathbb{E}_{\ell'_t\sim q_t}\left[\langle g(x_t), \ell'_t\rangle - \eta_3\langle g(x_t), \ell'_t\rangle^2\right] - \left(\langle g(x_t), \ell_t\rangle - \eta_3\langle g(x_t), \ell_t\rangle^2\right)\right.\right. \\
&\qquad\qquad\qquad \left.\left. - \eta_3 \sum_{t=1}^{n}\left(\mathbb{E}_{\ell_t\sim q_t}\left[\langle g(x_t), \ell_t\rangle^2\right] + \langle g(x_t), \ell_t\rangle^2\right)\right] - B_1(\ell_{1:n})\right] \\
=&\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathbb{E}_{\ell_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} \mathbb{E}_{\ell'_t\sim q_t}\left[H(\langle g(x_t), \ell'_t\rangle)\right] - H(\langle g(x_t), \ell_t\rangle)\right.\right. \\
&\qquad\qquad\qquad \left.\left. - \eta_3 \sum_{t=1}^{n}\left(\mathbb{E}_{\ell_t\sim q_t}\left[\langle g(x_t), \ell_t\rangle^2\right] + \langle g(x_t), \ell_t\rangle^2\right)\right] - B_1(\ell_{1:n})\right].
\end{aligned}$$

Using Jensen's inequality, this is upper bounded by

$$\begin{aligned}
&\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathbb{E}_{\ell_t,\ell'_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g\in\mathcal{G}}\left[\sum_{t=1}^{n} H(\langle g(x_t), \ell'_t\rangle) - H(\langle g(x_t), \ell_t\rangle)\right.\right. \\
&\qquad\qquad \left.\left. -\eta_3 \sum_{t=1}^{n}\left(\langle g(x_t), \ell'_t\rangle^2 + \langle g(x_t), \ell_t\rangle^2\right)\right] - B_1(\ell_{1:n})\right],
\end{aligned} \tag{11.9}$$

where $\ell'_{1:n}$ is a tangent sequence. We now claim that this is equal to

$$\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t,\ell'_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \Bigg[\sup_{g\in\mathcal{G}}\Bigg[\sum_{t=1}^{n} H(\langle g(x_t),\ell'_t\rangle) - H(\langle g(x_t),\ell_t\rangle)$$
$$- \eta_3 \sum_{t=1}^{n}\Big(\langle g(x_t),\ell'_t\rangle^2 + \langle g(x_t),\ell_t\rangle^2\Big)\Bigg] - \frac{1}{2}B_1(\ell_{1:n}) - \frac{1}{2}B_1(\ell'_{1:n})\Bigg].$$

This can be seen as follows: Let Q be the joint distribution over $\ell_1,\dots,\ell_n$ obtaining the supremum above, or if the supremum is not obtained let it be any point in a limit sequence approaching the supremum. Then the value of the B_1 term in (11.9) is equal to (respectively, ε-close to)

$$\begin{aligned}
\mathop{\mathbb{E}}_{Q} B_1(\ell_{1:n}) &= \eta_1 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{Q}\|\ell_t\|_1 + \eta_2 \sum_{t=1}^{n}\mathop{\mathbb{E}}_{Q}\|\ell_t\|_1^2 \\
&= \eta_1 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell_{1:t-1}} \mathbb{E}[\|\ell_t\|_1 \mid \ell_{1:t-1}] + \eta_2 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell_{1:t-1}} \mathbb{E}\big[\|\ell_t\|_1^2 \mid \ell_{1:t-1}\big] \\
&= \eta_1 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell_{1:t-1}} \mathbb{E}[\|\ell'_t\|_1 \mid \ell_{1:t-1}] + \eta_2 \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\ell_{1:t-1}} \mathbb{E}\Big[\|\ell'_t\|_1^2 \mid \ell_{1:t-1}\Big] \\
&= \mathop{\mathbb{E}}_{\ell_{1:n}} \mathop{\mathbb{E}}_{\ell'_{1:n}\mid\ell_{1:n}} B_1(\ell'_{1:n}).
\end{aligned}$$

Replacing ℓ_t with ℓ'_t follows from the definition of the tangent sequence, since ℓ'_t and ℓ_t are identically distributed, conditioned on $\ell_{1:t-1}$. This shows that we can replace $B_1(\ell_{1:n})$ with $B_1(\ell_{1:n})/2 + B_1(\ell'_{1:n})/2$ above, since we are working with the expectation.

We have now established that (11.9) is equal to

$$\left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \sup_{q_t\in\Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t,\ell'_t\sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \Bigg[\sup_{g\in\mathcal{G}}\Bigg[\sum_{t=1}^{n} \underbrace{H(\langle g(x_t),\ell'_t\rangle)}_{A_1} - \underbrace{H(\langle g(x_t),\ell_t\rangle)}_{A_2}$$
$$- \eta_3\Bigg(\sum_{t=1}^{n} \underbrace{\langle g(x_t),\ell'_t\rangle^2 + \langle g(x_t),\ell_t\rangle^2}_{A_3}\Bigg)\Bigg]$$
$$- \frac{\eta_1}{2}\Bigg(\sum_{t=1}^{n}\underbrace{\|\ell_t\|_1 + \|\ell'_t\|_1}_{A_4}\Bigg) - \frac{\eta_2}{2}\Bigg(\sum_{t=1}^{n}\underbrace{\|\ell_t\|_1^2 + \|\ell'_t\|_1^2}_{A_5}\Bigg)\Bigg].$$

Fix a time t and suppose the values of ℓ_t and ℓ'_t are exchanged. In this case the value of $A_1 - A_2$ is switched to $A_2 - A_1$, while the values of A_3, A_4, and A_5 are left unchanged. Appealing to Lemma 36, we can therefore introduce Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$ with equality as follows:

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{q_t \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t, \ell'_t \sim q_t} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H(\langle g(x_t), \ell'_t \rangle) - H(\langle g(x_t), \ell_t \rangle) \big) \right.\right.$$
$$\left. - \eta_3 \left(\sum_{t=1}^{n} \langle g(x_t), \ell'_t \rangle^2 + \langle g(x_t), \ell_t \rangle^2 \right) \right]$$
$$\left. - \frac{\eta_1}{2} \left(\sum_{t=1}^{n} \|\ell_t\|_1 + \|\ell'_t\|_1 \right) - \frac{\eta_2}{2} \left(\sum_{t=1}^{n} \|\ell_t\|_1^2 + \|\ell'_t\|_1^2 \right) \right].$$

Splitting the supremum, this is upper bounded by two times the following quantity

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{q_t \in \Delta(\mathcal{L})} \mathop{\mathbb{E}}_{\ell_t \sim q_t} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t H(\langle g(x_t), \ell_t \rangle) - \eta_3 \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle^2 \right] - \frac{\eta_1}{2} \sum_{t=1}^{n} \|\ell_t\|_1 - \frac{\eta_2}{2} \sum_{t=1}^{n} \|\ell_t\|_1^2 \right]$$
$$= \left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \sup_{\ell_t \in \mathcal{L}} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t H(\langle g(x_t), \ell_t \rangle) - \eta_3 \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle^2 \right] - \frac{\eta_1}{2} \sum_{t=1}^{n} \|\ell_t\|_1 - \frac{\eta_2}{2} \sum_{t=1}^{n} \|\ell_t\|_1^2 \right]$$
$$= \sup_{\boldsymbol{x}} \sup_{\boldsymbol{\ell}} \mathop{\mathbb{E}}_{\epsilon} \left[\sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - \eta_3 \sum_{t=1}^{n} \langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle^2 \right] \right.$$
$$\left. - \frac{\eta_1}{2} \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1 - \frac{\eta_2}{2} \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1^2 \right].$$

The first equality is somewhat subtle, but holds because at time n, the expression is linear in q_n so it is maximized at a point ℓ_n, allowing us to work backwards to remove the q_t distributions.

Introducing a coarse cover

We now break the process appearing in (11.8) into multiple terms, each of which will be handled by covering. Consider any fixed pair of trees $\boldsymbol{x}, \boldsymbol{\ell}$. Note that with the trees fixed (11.7) is at most

$$2 \cdot \mathop{\mathbb{E}}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - \eta_3 \sum_{t=1}^{n} \langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle^2 \right] - \mathop{\mathbb{E}}_{\epsilon} B_1(\boldsymbol{\ell}_{1:n}(\epsilon)).$$

We will focus on the supremum for now. We begin by adapting a trick from Rakhlin and Sridharan (2015) to introduce a coarse sequential cover at scale β. Let V' be a cover for $\mathcal{G}$

on the tree $\boldsymbol{x}$ with respect to L_∞/ℓ_∞ at scale $\beta/2$. Then the size of V' is $\mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, \boldsymbol{x})$, and

$$\max_{g\in\mathcal{G}} \max_{\epsilon\in\{\pm 1\}^n} \min_{\boldsymbol{v}'\in V'} \max_{t\in[n]} \|g(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{v}'_t(\epsilon)\|_\infty \leq \beta/2.$$

Recall that since $g(x) \in \mathbb{R}^K_+$ for all $g \in \mathcal{G}$, we may take each $\boldsymbol{v}' \in V'$ to have non-negative coordinates without loss of generality. Likewise, it follows that we may take each $\boldsymbol{v}' \in V'$ to have $\|\boldsymbol{v}'_t(\epsilon)\|_\infty \leq \sup_{x\in\mathcal{X}} \sup_{g\in\mathcal{G}} \|g(x)\|_\infty$ without loss of generality.

We construct a new β-cover V^1 from V' by defining for each tree $\boldsymbol{v}' \in V'$ a new tree $\boldsymbol{v}$ as follows:

$$\forall \epsilon \in \{\pm 1\}^n \; \forall t \in [n] \; \forall a \in [K]: \quad \boldsymbol{v}_t(\epsilon)_a = \max\{\boldsymbol{v}'_t(\epsilon)_a - \beta/2, 0\}.$$

It is easy to verify that for each time t and path ϵ we have $\|\boldsymbol{v}_t(\epsilon) - \boldsymbol{v}'_t(\epsilon)\|_\infty \leq \beta/2$, so V^1 is indeed a β-cover with respect to L_∞/ℓ_∞. More importantly, for each $g \in \mathcal{G}$ and path ϵ, there exists a tree $\boldsymbol{v} \in V'$ that is β-close in the L_∞/ℓ_∞ sense and has $\boldsymbol{v}_t(\epsilon)_a \leq g(\boldsymbol{x}_t(\epsilon))_a$ coordinate-wise. We will let $\boldsymbol{v}^1[\epsilon, g]$ denote this tree, and it is constructed by taking the $\beta/2$-close tree $\boldsymbol{v}'$ promised by the definition of V', then performing the clipping operation above to get the corresponding β-close element of V^1. The clipping operation and $\beta/2$ closeness of $\boldsymbol{v}'$ imply that for each time $t \in [n]$ and coordinate $a \in [K]$,

$$\begin{aligned}
\boldsymbol{v}^1_t[\epsilon, g]_a - g(\boldsymbol{x}_t(\epsilon))_a &= \max\{\boldsymbol{v}'_t(\epsilon)_a - \beta/2, 0\} - g(\boldsymbol{x}_t(\epsilon))_a \\
&\leq \max\{\|\boldsymbol{v}'_t(\epsilon) - g(\boldsymbol{x}_t(\epsilon))\|_\infty + g(\boldsymbol{x}_t(\epsilon))_a - \beta/2, 0\} - g(\boldsymbol{x}_t(\epsilon))_a \\
&\leq \max\{g(\boldsymbol{x}_t(\epsilon))_a, 0\} - g(\boldsymbol{x}_t(\epsilon))_a = 0.
\end{aligned}$$

This establishes the desired ordering on coordinates. Returning to the process at hand, we have

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{g\in\mathcal{G}} \left[\sum_{t=1}^n \epsilon_t H\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon)\rangle) - \eta_3 \sum_{t=1}^n \langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon)\rangle^2 \right].$$

Now we add and subtract terms involving the covering element $\boldsymbol{v}^1(\epsilon, g)$:

$$= \mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \Bigg[\sum_{t=1}^{n} \epsilon_t H(\big\langle \boldsymbol{v}_t^1[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle) - \eta_3 \sum_{t=1}^{n} \langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle^2 + \sum_{t=1}^{n} \epsilon_t H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - \epsilon_t H(\big\langle \boldsymbol{v}_t^1[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle) \Bigg].$$

We now invoke the coordinate domination property of $\boldsymbol{v}^1[\epsilon, g]$ described above. Observe that since $g(\boldsymbol{x}_t(\epsilon))$, $\boldsymbol{v}_t^1[\epsilon, g]$, and $\boldsymbol{\ell}_t(\epsilon)$ are all nonnegative coordinate-wise, it holds that $\langle \boldsymbol{v}_t^1[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \rangle^2 \leq \langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle^2$. Consequently, we can replace the offset term (not involving ϵ_t) with a similar term involving $\boldsymbol{v}_t^1[\epsilon, g]$

$$\leq \mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \Bigg[\sum_{t=1}^{n} \epsilon_t H(\big\langle \boldsymbol{v}_t^1[\epsilon, f], \boldsymbol{\ell}_t(\epsilon) \big\rangle) - \eta_3 \sum_{t=1}^{n} \big\langle \boldsymbol{v}_t^1[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle^2 + \sum_{t=1}^{n} \epsilon_t H(\langle f(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - \epsilon_t H(\big\langle \boldsymbol{v}_t^1[\epsilon, f], \boldsymbol{\ell}_t(\epsilon) \big\rangle) \Bigg].$$

Splitting the supremum and gathering terms, this implies that $\mathcal{V}_n^{\mathrm{ol}}$ is upper bounded by

$$\underbrace{\mathbb{E}_{\epsilon} \sup_{\boldsymbol{v}^1 \in V^1} \left[\sum_{t=1}^{n} \epsilon_t H(\big\langle \boldsymbol{v}_t^1(\epsilon), \boldsymbol{\ell}_t(\epsilon) \big\rangle) - \eta_3 \sum_{t=1}^{n} \big\langle \boldsymbol{v}_t^1(\epsilon), \boldsymbol{\ell}_t(\epsilon) \big\rangle^2 \right]}_{(\star)}$$

$$+ \underbrace{\mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - \epsilon_t H(\big\langle \boldsymbol{v}_t^1[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle) \right] - \mathbb{E}_{\epsilon} B_1(\boldsymbol{\ell}_{1:n}(\epsilon))}_{(\star\star)}.$$

Bounding $(\star)$

We appeal to Lemma 34 with a class of real-valued trees

$$U := \Big\{ \epsilon \mapsto \big(\big\langle \boldsymbol{v}_t^1(\epsilon), \boldsymbol{\ell}_t(\epsilon) \big\rangle \big)_{t \leq n} \mid \boldsymbol{v}^1 \in V^1 \Big\}.$$

The class U has range contained in $[-1, +1]$, since $|\langle \boldsymbol{v}_t^1(\epsilon), \boldsymbol{\ell}_t(\epsilon) \rangle| \leq \|\boldsymbol{v}_t^1(\epsilon)\|_{\infty} \|\boldsymbol{\ell}_t(\epsilon)\|_1 \leq 1$, where these norm bounds are by assumption on $\mathcal{G}$ and $\mathcal{L}$. Recall that $H(x) = x - \eta_3 x^2$. We therefore conclude that

$$
\begin{aligned}
(\star) &= \mathbb{E}_{\epsilon} \sup_{\boldsymbol{v}^1 \in V^1} \left[\sum_{t=1}^{n} \epsilon_t H\big(\big\langle \boldsymbol{v}_t^1(\epsilon), \boldsymbol{\ell}_t(\epsilon) \big\rangle\big) - \eta_3 \sum_{t=1}^{n} \big\langle \boldsymbol{v}_t^1(\epsilon), \boldsymbol{\ell}_t(\epsilon) \big\rangle^2 \right] \\
&\leq 2\frac{1+\eta_3^2}{\eta_3} \log\big|V^1\big| = 2\frac{1+\eta_3^2}{\eta_3} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, \boldsymbol{x}).
\end{aligned}
$$

Bounding $(\star\star)$

Fix $\alpha > 0$ and let $N = \lfloor \log(\beta/\alpha) \rfloor - 1$. For each $i \geq 1$ define $\varepsilon_i = \beta e^{-(i-1)}$, and for each $i > 1$ let V^i be a sequential cover of $\mathcal{G}$ on $\boldsymbol{x}$ at scale ε_i with respect to L_∞/ℓ_∞ (keeping in mind that V^1 is defined as in the preceding section). For a given path $\epsilon \in \{\pm 1\}^n$ and $g \in \mathcal{G}$, let $\boldsymbol{v}^i[\epsilon, g]$ denote the ε_i-close element of V^i. Below, we will only evaluate $H(x) = x - \eta_3 x^2$ over the domain $[-1, +1]$; it is $(1 + 2\eta_3)$-Lipschitz over this domain. Then the leading term of $(\star\star)$ is equal to

$$
\mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - H\big(\big\langle \boldsymbol{v}_t^1[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big)\big) \right].
$$

Introducing the covering elements defined above to this expression, we have the equality

$$
\begin{aligned}
&= \mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - H\big(\big\langle \boldsymbol{v}_t^N[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big)\big) \right. \\
&\qquad \left. + \sum_{i=1}^{N-1} \sum_{t=1}^{n} \epsilon_t \big(H\big(\big\langle \boldsymbol{v}_t^{i+1}[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big) - H\big(\big\langle \boldsymbol{v}_t^i[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big)\big) \right] \\
&\leq \underbrace{\mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - H\big(\big\langle \boldsymbol{v}_t^N[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big)\big) \right]}_{\triangleq C_N} \\
&\qquad + \sum_{i=1}^{N-1} \underbrace{\mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H\big(\big\langle \boldsymbol{v}_t^{i+1}[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big) - H\big(\big\langle \boldsymbol{v}_t^i[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle\big)\big) \right]}_{\triangleq C_i}.
\end{aligned}
$$

Bounding C_N

We first bound C_N in terms of one of the terms appearing in B_1.

$$\begin{aligned}
C_N =& \mathop{\mathbb{E}}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - H\big(\big\langle \boldsymbol{v}_t^N[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle \big) \big) \right] \\
\leq& \mathop{\mathbb{E}}_{\epsilon} \left[\sum_{t=1}^{n} \sup_{g \in \mathcal{G}} \big| H(\langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle) - H\big(\big\langle \boldsymbol{v}_t^N[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle \big) \big| \right] \\
\leq& (1 + 2\eta_3) \mathop{\mathbb{E}}_{\epsilon} \left[\sum_{t=1}^{n} \sup_{g \in \mathcal{G}} \big| \langle g(\boldsymbol{x}_t(\epsilon)), \boldsymbol{\ell}_t(\epsilon) \rangle - \big\langle \boldsymbol{v}_t^N[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle \big| \right].
\end{aligned}$$

To proceed, we apply Hölder's inequality.

$$\begin{aligned}
&\leq (1 + 2\eta_3) \mathop{\mathbb{E}}_{\epsilon} \left[\sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1 \sup_{g \in \mathcal{G}} \big\| g(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{v}_t^N[\epsilon, g] \big\|_\infty \right] \\
&\leq (1 + 2\eta_3) \max_{\epsilon} \sup_{g \in \mathcal{G}} \max_{t \in [n]} \big\| g(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{v}_t^N[\epsilon, g] \big\|_\infty \cdot \mathop{\mathbb{E}}_{\epsilon} \left[\sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1 \right] \\
&\leq (1 + 2\eta_3) e^2 \alpha \cdot \mathop{\mathbb{E}}_{\epsilon} \left[\sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1 \right].
\end{aligned}$$

The first inequality uses that $\epsilon_t \in \{\pm 1\}$, while the second uses the Lipschitzness of H over $[-1, +1]$. The third and fourth are both applications of Hölder's inequality, first to the ℓ_1/ℓ_∞ dual pairing, and then to for the distributions over L_1/L_∞. Finally, the definition of the covering element $\boldsymbol{v}_t^N$—in particular, that it is an L_∞/ℓ_∞-cover—implies that the supremum term is bounded by $\varepsilon_N \leq e^2 \cdot \alpha$, which yields the final bound.

Bounding C_i

Our goal is to bound

$$C_i = \mathop{\mathbb{E}}_{\epsilon} \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_t \big(H\big(\big\langle \boldsymbol{v}_t^{i+1}[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle \big) - H\big(\big\langle \boldsymbol{v}_t^{i}[\epsilon, g], \boldsymbol{\ell}_t(\epsilon) \big\rangle \big) \big) \right].$$

We define a class W of real-valued trees as follows. Let $1 \leq a \leq |V^i|$ and $1 \leq b \leq |V^{i+1}|$, and fix an arbitrary ordering $\boldsymbol{v}^a \in V^i$ and $\boldsymbol{v}^b \in V^{i+1}$ of the elements of V^i/V^{i+1}. For each pair (a, b) define a tree $\boldsymbol{w}^{(a,b)}$ via

$$\boldsymbol{w}_t^{(a,b)}(\epsilon) = \begin{cases} H\big(\big\langle \boldsymbol{v}_t^b(\epsilon), \boldsymbol{\ell}_t(\epsilon)\big\rangle\big) - H(\langle \boldsymbol{v}_t^a(\epsilon), \boldsymbol{\ell}_t(\epsilon)\rangle), & \exists g \in \mathcal{G} \text{ s.t. } \boldsymbol{v}^a = \boldsymbol{v}[\epsilon, g]^i, \boldsymbol{v}^b = \boldsymbol{v}[\epsilon, g]^{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

Then C_i is bounded by

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{w} \in W} \sum_{t=1}^n \epsilon_t \boldsymbol{w}_t(\epsilon).$$

Then Lemma 35 implies that for any fixed $\eta > 0$,

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{w} \in W} \left[\sum_{t=1}^n \epsilon_t \boldsymbol{w}_t(\epsilon) - \eta \boldsymbol{w}_t^2(\epsilon)\right] \leq \frac{\log|W|}{2\eta}.$$

Rearranging and applying subadditivity of the supremum, this implies

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{w} \in W} \sum_{t=1}^n \epsilon_t \boldsymbol{w}_t(\epsilon) \leq \eta \cdot \mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{w} \in W} \sum_{t=1}^n \boldsymbol{w}_t^2(\epsilon) + \frac{\log|W|}{2\eta}.$$

Optimizing over η (which is admissible because the statement above is a deterministic inequality) leads to a further bound of

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{w} \in W} \sum_{t=1}^n \epsilon_t \boldsymbol{w}_t(\epsilon) \leq \sqrt{2 \mathop{\mathbb{E}}_{\epsilon} \sup_{\boldsymbol{w} \in W} \sum_{t=1}^n \boldsymbol{w}_t^2(\epsilon) \cdot \log|W|}.$$

We proceed to bound each term in the square root. For the logarithmic term, by construction we have $|W| \leq |V^i||V^{i+1}| \leq |V^{i+1}|^2 = \mathcal{N}_{\infty,\infty}(\varepsilon_{i+1}, \mathcal{G}, \boldsymbol{x})^2$.

For the variance, let $\boldsymbol{w}^{(a,b)} \in W$ and the path ϵ be fixed. There are two cases: Either $\boldsymbol{w}(\epsilon) = \boldsymbol{0}$, or there exists $g \in \mathcal{G}$, such that $\boldsymbol{v}^a = \boldsymbol{v}[\epsilon, g]^i$ and $\boldsymbol{v}^b = \boldsymbol{v}[\epsilon, g]^{i+1}$. The former case is trivial while for the latter, in a similar way to the bound for C_N, we get

$$\begin{aligned}
\sum_{t=1}^n \boldsymbol{w}_t^{(a,b)}(\epsilon)^2 &= \sum_{t=1}^n \Big(H\big(\big\langle \boldsymbol{v}_t^{i+1}[\epsilon, g], \boldsymbol{\ell}_t(\epsilon)\big\rangle\big) - H\big(\big\langle \boldsymbol{v}_t^i[\epsilon, g], \boldsymbol{\ell}_t(\epsilon)\big\rangle\big)\Big)^2 \\
&\leq (1 + 2\eta_3)^2 \sum_{t=1}^n \Big(\big\langle \boldsymbol{v}_t^{i+1}[\epsilon, g], \boldsymbol{\ell}_t(\epsilon)\big\rangle - \big\langle \boldsymbol{v}_t^i[\epsilon, g], \boldsymbol{\ell}_t(\epsilon)\big\rangle\Big)^2 \\
&\leq (1 + 2\eta_3)^2 \sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1^2 \big\|\boldsymbol{v}_t^{i+1}[\epsilon, g] - \boldsymbol{v}_t^i[\epsilon, g]\big\|_\infty^2 \\
&\leq (1 + 2\eta_3)^2 \max_{\epsilon'} \max_{t \in [n]} \big\|\boldsymbol{v}_t^{i+1}[\epsilon', g] - \boldsymbol{v}_t^i[\epsilon', g]\big\|_\infty^2 \cdot \sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1^2.
\end{aligned}$$

Where we have used Lipschitzness of H in the first inequality and Hölder's inequality in the second and third.

Finally, using the L_∞/ℓ_∞ cover property of $\boldsymbol{v}^i[\epsilon, g]$ and $\boldsymbol{v}^{i+1}[\epsilon, g]$ and the triangle inequality, we have

$$\begin{aligned}
&\max_{\epsilon}\max_{t\in[n]}\left\|\boldsymbol{v}_t^{i+1}[\epsilon, g] - \boldsymbol{v}_t^{i}[\epsilon, g]\right\|_\infty \\
&\leq \max_{\epsilon}\max_{t\in[n]}\left\|\boldsymbol{v}_t^{i+1}[\epsilon, g] - g(\boldsymbol{x}_t(\epsilon))\right\|_\infty + \max_{\epsilon}\max_{t\in[n]}\left\|g(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{v}_t^{i}[\epsilon, g]\right\|_\infty \\
&\leq \varepsilon_i + \varepsilon_{i+1} \leq 2\varepsilon_i.
\end{aligned}$$

We have just shown that for every sequence ϵ and every $\boldsymbol{w}^{(a,b)} \in W$, $\sum_{t=1}^n \boldsymbol{w}_t^{(a,b)}(\epsilon)^2 \leq 4(1+2\eta_3)^2\varepsilon_i^2 \cdot \sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1^2$. It follows that

$$\mathop{\mathbb{E}}_{\epsilon}\sup_{\boldsymbol{w}\in W}\sum_{t=1}^n \boldsymbol{w}_t(\epsilon)^2 \leq 4(1+2\eta_3)^2\varepsilon_i^2 \cdot \mathop{\mathbb{E}}_{\epsilon}\sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1^2.$$

Plugging this bound back into the main inequality, we have shown

$$\mathop{\mathbb{E}}_{\epsilon}\sup_{\boldsymbol{w}\in W}\sum_{t=1}^n \epsilon_t \boldsymbol{w}_t(\epsilon) \leq 4e(1+2\eta_3)\varepsilon_{i+1}\sqrt{\mathop{\mathbb{E}}_{\epsilon}\sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1^2 \cdot \log\mathcal{N}_{\infty,\infty}(\varepsilon_{i+1}, \mathcal{G}, \boldsymbol{x})}.$$

Final bound on $(\star\star)$

Collecting terms, we have shown that

$$\begin{aligned}
&(\star\star) \\
&\leq (1+2\eta_3)e^2\alpha \cdot \mathop{\mathbb{E}}_{\epsilon}\left[\sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1\right] \\
&\quad + 4e(1+2\eta_3)\sqrt{\mathop{\mathbb{E}}_{\epsilon}\sum_{t=1}^n \|\boldsymbol{\ell}_t(\epsilon)\|_1^2}\sum_{i=1}^{N-1}\varepsilon_{i+1}\sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon_{i+1}, \mathcal{G}, \boldsymbol{x})} - \mathop{\mathbb{E}}_{\epsilon} B_1(\boldsymbol{\ell}_{1:n}(\epsilon)).
\end{aligned} \tag{11.10}$$

Following the standard Dudley chaining proof, we have

$$\sum_{i=1}^{N-1} \varepsilon_{i+1} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon_{i+1}, \mathcal{G}, \boldsymbol{x})} 2 \leq \sum_{i=1}^{N} (\varepsilon_i - \varepsilon_{i+1}) \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon_i, \mathcal{G}, \boldsymbol{x})}$$
$$\leq 2 \int_{\varepsilon_{N+1}}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, \boldsymbol{x})} d\varepsilon.$$

We further upper bound by

$$\leq 2 \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, \boldsymbol{x})} d\varepsilon \leq 2 \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon.$$

We are using the definition of N, which implies that $\alpha \leq \varepsilon_{N+1}$.

Now recall the definition of $B_1(\boldsymbol{\ell}_{1:n}(\epsilon))$:

$$B_1(\boldsymbol{\ell}_{1:n}(\epsilon)) = \eta_1 \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1 + \eta_2 \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1^2.$$

Taking $\eta_1 \geq (1 + 2\eta_3)e^2\alpha$, the first term in B_1 cancels out the first term in (11.10), leaving us with

$$(\star\star) \leq 4e(1 + 2\eta_3) \sqrt{\mathop{\mathbb{E}}_{\epsilon} \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1^2} \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon - \eta_2 \mathop{\mathbb{E}}_{\epsilon} \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1^2$$
$$\leq 4e(1 + 2\eta_3) \left(\frac{\eta_4}{4} \mathop{\mathbb{E}}_{\epsilon} \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1^2 + \frac{1}{\eta_4} \right) \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon - \eta_2 \mathop{\mathbb{E}}_{\epsilon} \sum_{t=1}^{n} \|\boldsymbol{\ell}_t(\epsilon)\|_1^2.$$

Where the last step applies for any $\eta_4 > 0$ by the AM-GM inequality. For any $\eta_2 \geq e(1 + 2\eta_3)\eta_4$, the first and third terms cancel, leaving us with an upper bound of

$$(\star\star) \leq \frac{4e(1 + 2\eta_3)}{\eta_4} \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon.$$

This term does not depend on the trees $\boldsymbol{x}$ or $\boldsymbol{\ell}$, so we are done with $(\star\star)$.

Final bound

Under the assumptions on $\eta_1, \eta_2, \eta_3, \eta_4, \alpha$, and β, the bounds on $(\star)$ and $(\star\star)$ we have established imply

$$\mathcal{V}_n^{\mathrm{ol}} \le 2\frac{1+\eta_3^2}{\eta_3} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) + \frac{4e(1+2\eta_3)}{\eta_4} \int_\alpha^\beta \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon.$$

The definition of $\mathcal{V}_n^{\mathrm{ol}}$ implies that there exists an algorithm with regret bounded by $\mathcal{V}_n^{\mathrm{ol}} + B(p_{1:n}, \ell_{1:n})$ on every sequence. The final regret inequality is

$$\begin{aligned}
&\sum_{t=1}^n \mathop{\mathbb{E}}_{s\sim p_t} \langle s, \ell_t \rangle - \inf_{g\in\mathcal{G}} \sum_{t=1}^n \langle f(x_t), \ell_t \rangle \\
&\le 2\eta_3 \sum_{t=1}^n \mathop{\mathbb{E}}_{s\sim p_t} \langle s, \ell_t \rangle^2 + 2\frac{1+\eta_3^2}{\eta_3} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) \\
&\qquad + 4e(1+2\eta_3)\left(\frac{\eta_4}{4}\sum_{t=1}^n \|\ell_t\|_1^2 + \frac{1}{\eta_4}\right) \int_\alpha^\beta \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon + (1+2\eta_3)e^2\alpha \sum_{t=1}^n \|\ell_t\|_1.
\end{aligned}$$

To obtain the bound in the theorem statement, we rebind $\eta = \eta_3, \lambda = \eta_4$ and use the assumption $\eta \le 1$.

11.5.5 Proofs for Remaining Minimax Results

Our bandit results require a generalization of Theorem 44 to the case where losses and the class $\mathcal{G}$ may not be bounded by 1.

Corollary 15. Suppose we are in the setting of Theorem 44, but with the bounds $\|\ell\|_1 \le R$ for all $\ell \in \mathcal{L}$ and $\|s\|_\infty \le B$ for all $s \in \mathcal{S}$. For any constants $\eta \in (0, 1]$, $\lambda > 0$, and $\beta > \alpha > 0$, there exists an algorithm making predictions in $\mathcal{S}$ that attains a regret guarantee of

$$\begin{aligned}
&\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle - \inf_{g\in\mathcal{G}}\sum_{t=1}^{n}\langle g(x_t), \ell_t\rangle \\
&\leq \frac{2\eta}{RB}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle^2 + \frac{4RB}{\eta}\log\mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) + 3e^2\alpha\sum_{t=1}^{n}\|\ell_t\|_1 \\
&\quad + 12e\left(\frac{\lambda}{4R}\sum_{t=1}^{n}\|\ell_t\|_1^2 + \frac{R}{\lambda}\right)\int_{\alpha}^{\beta}\sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)}d\varepsilon.
\end{aligned}$$

Furthermore, if upper bounds $\sum_{t=1}^{n}\|\ell_t\|_1^2 \leq V$ and $\sum_{t=1}^{n}\mathbb{E}_{s_t\sim p_t}\langle s_t, \ell_t\rangle^2 \leq \widetilde{V}$ are known in advance, η and λ can be selected to guarantee regret

$$\begin{aligned}
&\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle - \inf_{g\in\mathcal{G}}\sum_{t=1}^{n}\langle g(x_t), \ell_t\rangle \\
&\leq 8\sqrt{\widetilde{V}\cdot\log\mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n)} + 8RB\log\mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) \\
&\qquad + 12e\sqrt{V}\int_{\alpha}^{\beta}\sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)}d\varepsilon + 3e\alpha\sum_{t=1}^{n}\|\ell_t\|_1.
\end{aligned}$$

Proof of Corollary 15. Apply Theorem 44 with losses ℓ_t/R and class $\mathcal{G}/B$. The preconditions of the theorem are satisified, so it implies existence of an algorithm making predictions in $\mathcal{S}/B$ with regret bound

$$\begin{aligned}
&\frac{1}{R}\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle - \frac{1}{R}\inf_{g'\in\mathcal{G}/B}\sum_{t=1}^{n}\langle g'(x_t), \ell_t\rangle \\
&\leq \frac{2\eta}{R^2}\sum_{t=1}^{n}\mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle^2 + \frac{4}{\eta}\log\mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}/B, n) + \frac{3e^2\alpha}{R}\sum_{t=1}^{n}\|\ell_t\|_1 \\
&\quad + 12e\left(\frac{\lambda}{4R^2}\sum_{t=1}^{n}\|\ell_t\|_1^2 + \frac{1}{\lambda}\right)\int_{\alpha}^{\beta}\sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}/B, n)}d\varepsilon.
\end{aligned}$$

Rescaling both sides by BR and letting $\hat{s}_t = s_t \cdot B$ (so $\hat{s}_t \in \mathcal{S}$), this implies

$$\begin{aligned}
&\sum_{t=1}^{n} \mathop{\mathbb{E}}_{\hat{s}_t \sim p_t} \langle \hat{s}_t, \ell_t \rangle - \inf_{g \in \mathcal{G}} \sum_{t=1}^{n} \langle g(x_t), \ell_t \rangle \\
&\leq \frac{2\eta}{RB} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\hat{s}_t \sim p_t} \langle \hat{s}_t, \ell_t \rangle^2 + \frac{4RB}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}/B, n) + 3e^2 \alpha B \sum_{t=1}^{n} \|\ell_t\|_1 \\
&\quad + 12e \left(\frac{\lambda B}{4R} \sum_{t=1}^{n} \|\ell_t\|_1^2 + \frac{RB}{\lambda} \right) \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}/B, n)} d\varepsilon. \\
&\leq \frac{2\eta}{RB} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\hat{s}_t \sim p_t} \langle \hat{s}_t, \ell_t \rangle^2 + \frac{4RB}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta B/2, \mathcal{G}, n) + 3e^2 \alpha B \sum_{t=1}^{n} \|\ell_t\|_1 \\
&\quad + 12e \left(\frac{\lambda B}{4R} \sum_{t=1}^{n} \|\ell_t\|_1^2 + \frac{RB}{\lambda} \right) \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon B, \mathcal{G}, n)} d\varepsilon.
\end{aligned}$$

Using a change of variables in the Dudley integral, we get

$$\begin{aligned}
&\leq \frac{2\eta}{RB} \sum_{t=1}^{n} \mathop{\mathbb{E}}_{\hat{s}_t \sim p_t} \langle \hat{s}_t, \ell_t \rangle^2 + \frac{4RB}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta B/2, \mathcal{G}, n) + 3e^2 \alpha B \sum_{t=1}^{n} \|\ell_t\|_1 \\
&\quad + 12e \left(\frac{\lambda}{4R} \sum_{t=1}^{n} \|\ell_t\|_1^2 + \frac{R}{\lambda} \right) \int_{\alpha B}^{\beta B} \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{G}, n)} d\varepsilon.
\end{aligned}$$

The final result follows by rebinding $\alpha' = \alpha B$ and $\beta' = \beta B$.

For the second claim, apply the upper bounds to obtain

$$\begin{aligned}
&\frac{2\eta}{RB} \tilde{V} + \frac{4RB}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n) + 3e^2 \alpha B \sum_{t=1}^{n} \|\ell_t\|_1 \\
&+ 12e \left(\frac{\lambda}{4R} V + \frac{R}{\lambda} \right) \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{G}, n)} d\epsilon.
\end{aligned}$$

Now set $\lambda = 2R/\sqrt{V}$ and $\eta = \sqrt{2}RB\sqrt{\log \mathcal{N}_{\infty,\infty}(\beta/2, \mathcal{G}, n)/\tilde{V}} \wedge 1$ to obtain the claimed bound. Note that the range term arises from the constraint that $\eta \in (0, 1]$. $\square$

Proof of Theorem 41. Recall that we use the reduction:

- Initialize full information algorithm whose existence is guaranteed by Theorem 44 with $\mathcal{G} = \phi^{\gamma} \circ \mathcal{F}$:

- For time $t = 1, \dots, n$:
  - Receive x_t and define $P_t(a) \triangleq \mathbb{E}_{s_t \sim p_t} \frac{s_t(a)}{\sum_{a' \in [K]} s_t(a')}$, where p_t is the output of the full information algorithm at time t.
  - Sample action $a_t \sim P_t^\mu$ and feed importance-weighted loss $\hat{\ell}_t(a) = \mathbf{1}\{a_t = a\}\ell_t(a)/P_t^\mu(a)$ into the full information algorithm.

With this setup, Corollary 15 guarantees that the following deterministic regret inequality holds for every sequence of outcomes (i.e. for every sequence $a_1, \dots, a_n$ sampled by the algorithm):

$$\begin{aligned}
&\sum_{t=1}^n \mathop{\mathbb{E}}_{s_t \sim p_t} \left\langle s_t, \hat{\ell}_t \right\rangle - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \left\langle \phi^\gamma(f(x_t)), \hat{\ell}_t \right\rangle \\
&\leq \frac{2\eta}{RB} \sum_{t=1}^n \mathop{\mathbb{E}}_{s_t \sim p_t} \left\langle s_t, \hat{\ell}_t \right\rangle^2 + \frac{4RB}{\eta} \log \mathcal{N}_{\infty,\infty}(\beta/2, \phi^\gamma \circ \mathcal{F}, n) + 3e^2\alpha \sum_{t=1}^n \left\|\hat{\ell}_t\right\|_1 \\
&\quad + 12e\left(\frac{\lambda}{4R} \sum_{t=1}^n \left\|\hat{\ell}_t\right\|_1^2 + \frac{R}{\lambda}\right) \int_\alpha^\beta \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \phi^\gamma \circ \mathcal{F}, n)} d\varepsilon,
\end{aligned}$$

where the boundedness of the ramp loss implies $B \leq 1$ and the smoothing factor μ in P_t^μ guarantees $R \leq 1/\mu$. Taking expectation over the draw of $a_1, \dots, a_n$, for any fixed $f \in \mathcal{F}$ we obtain the inequality

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^n \mathop{\mathbb{E}}_{s_t \sim p_t} \left\langle s_t, \hat{\ell}_t \right\rangle - \sum_{t=1}^n \left\langle \phi^\gamma(f(x_t)), \hat{\ell}_t \right\rangle\right] \\
&\leq \mathbb{E}\left[\frac{2\eta}{1/\mu} \sum_{t=1}^n \mathbb{E}\left[\mathop{\mathbb{E}}_{s_t \sim p_t} \left\langle s_t, \hat{\ell}_t \right\rangle^2 \mid \mathcal{J}_t\right] + \frac{4}{\eta\mu} \log \mathcal{N}_{\infty,\infty}(\beta/2, \phi^\gamma \circ \mathcal{F}, n) + 3e^2\alpha \sum_{t=1}^n \mathbb{E}\left[\left\|\hat{\ell}_t\right\|_1 \mid \mathcal{J}_t\right] \right. \\
&\quad \left. + 12e\left(\frac{\lambda}{4/\mu} \sum_{t=1}^n \mathbb{E}\left[\left\|\hat{\ell}_t\right\|_1^2 \mid \mathcal{J}_t\right] + \frac{1}{\lambda\mu}\right) \int_\alpha^\beta \sqrt{\log \mathcal{N}_{\infty,\infty}(\varepsilon, \phi^\gamma \circ \mathcal{F}, n)} d\varepsilon\right],
\end{aligned}$$

where the filtration $\mathcal{J}_t$ is defined as in Lemma 32. Using that the importance weighted losses are unbiased, the left-hand side is equal to

$$\mathbb{E}\left[\sum_{t=1}^n \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s_t, \ell_t \rangle - \sum_{t=1}^n \langle \phi^\gamma(f(x_t)), \ell_t \rangle\right].$$

We also have the following three properties, where the first two use that $\hat{\ell}_t$ is 1-sparse, and the last follows from Lemma 32:

1. $\mathbb{E}\left[\left\|\hat{\ell}_t\right\|_1 \mid \mathcal{J}_t\right] = \sum_{a\in[K]} P_t^\mu(a)\hat{\ell}_t(a) = \sum_{a\in[K]} \ell_t(a) \le K.$

2. $\mathbb{E}\left[\left\|\hat{\ell}_t\right\|_1^2 \mid \mathcal{J}_t\right] = \sum_{a\in[K]} P_t^\mu(a)\hat{\ell}_t^2(a) = \sum_{a\in[K]} \frac{\ell_t(a)}{P_t^\mu(a)} \le \frac{K}{\mu}.$

3. $\mathbb{E}\left[\mathbb{E}_{s_t\sim p_t}\left\langle s_t, \hat{\ell}_t\right\rangle^2 \mid \mathcal{J}_t\right] \le K^2.$

Together, these facts yield the bound

$$\mathbb{E}\left[\sum_{t=1}^n \mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle - \sum_{t=1}^n \langle \phi^\gamma(f(x_t)), \ell_t\rangle\right] \le \frac{2\eta}{1/\mu}K^2 n + \frac{4}{\eta\mu}\log\mathcal{N}_{\infty,\infty}(\beta/2, \phi^\gamma\circ\mathcal{F}, n) + 3e^2\alpha K n$$
$$+ 12e\left(\frac{\lambda K n}{4} + \frac{1}{\lambda\mu}\right)\int_\alpha^\beta \sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon, \phi^\gamma\circ\mathcal{F}, n)}d\varepsilon.$$

Optimizing η and λ (as in the proof of the second claim of Corollary 15) leads to a bound of

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^n \mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle - \sum_{t=1}^n \langle \phi^\gamma(f(x_t)), \ell_t\rangle\right] \\
&\le 4\sqrt{2K^2 n\log\mathcal{N}_{\infty,\infty}(\beta/2, \phi^\gamma\circ\mathcal{F}, n)} + \frac{8}{\mu}\log\mathcal{N}_{\infty,\infty}(\beta/2, \phi^\gamma\circ\mathcal{F}, n) \\
&\quad + 3e^2\alpha K n + 12e\sqrt{\frac{Kn}{\mu}}\int_\alpha^\beta \sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon, \phi^\gamma\circ\mathcal{F}, n)}d\varepsilon.
\end{aligned}$$

Since ϕ^γ is $\frac{1}{\gamma}$-Lipschitz with respect to the ℓ_∞ norm (as a coordinate-wise mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$), we can upper bound in terms of the covering numbers for the original class:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^n \mathop{\mathbb{E}}_{s_t\sim p_t}\langle s_t, \ell_t\rangle - \sum_{t=1}^n \langle \phi^\gamma(f(x_t)), \ell_t\rangle\right] \\
&\le 4\sqrt{2K^2 n\log\mathcal{N}_{\infty,\infty}(\gamma\beta/2, \mathcal{F}, n)} + \frac{8}{\mu}\log\mathcal{N}_{\infty,\infty}(\gamma\beta/2, \mathcal{F}, n) \\
&\quad + 3e^2\alpha K n + 12e\sqrt{\frac{Kn}{\mu}}\int_\alpha^\beta \sqrt{\log\mathcal{N}_{\infty,\infty}(\gamma\varepsilon, \mathcal{F}, n)}d\varepsilon.
\end{aligned}$$

Using a change of variables and the reparameterization $\alpha' = \alpha\gamma$, $\beta' = \beta\gamma$, the right hand side equals

$$4\sqrt{2K^2 n\log\mathcal{N}_{\infty,\infty}(\beta'/2, \mathcal{F}, n)} + \frac{8}{\mu}\log\mathcal{N}_{\infty,\infty}(\beta'/2, \mathcal{F}, n)$$
$$+ \frac{1}{\gamma}\left(3e^2\alpha K n + 12e\sqrt{\frac{Kn}{\mu}}\int_{\alpha'}^{\beta'} \sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n)}d\varepsilon\right).$$

Lastly, via Lemma 31, we have

$$\sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \langle s_t, \ell_t \rangle \geq \sum_{t=1}^{n} \mathop{\mathbb{E}}_{s_t \sim p_t} \frac{\sum_{a \in [K]} s_t(a) \ell_t(a)}{\sum_{a \in [K]} s_t(a)} = \sum_{t=1}^{n} \mathop{\mathbb{E}}_{a_t \sim P_t} \ell_t(a_t).$$

Finally, the definition of the smoothed distribution P_t^{μ} and boundedness of ℓ immediately implies

$$\sum_{t=1}^{n} \mathop{\mathbb{E}}_{a_t \sim P_t} \ell_t(a_t) \geq \sum_{t=1}^{n} \mathop{\mathbb{E}}_{a_t \sim P_t^{\mu}} \ell_t(a_t) - \mu K n. \qquad \square$$

Proof of Proposition 19. Suppose $\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n) \propto \varepsilon^{-p}$.

- When $p \geq 2$, it suffices to set $\beta = \mathrm{rad}_{\infty,\infty}(\mathcal{F}, n)$, $\mu = (Kn)^{-1/(p+1)} \gamma^{-p/(p+1)}$, and $\alpha = 1/(Kn\mu)^{1/p}$ in Theorem 41 to obtain $\widetilde{O}\big((Kn/\gamma)^{p/(p+1)}\big)$.
- When $p \in (0, 2]$, it suffices to set $\alpha = 1/(Kn)$, $\mu = (Kn)^{-2/(p+4)} \gamma^{-2p/(4+p)}$, and $\beta = \gamma^{2/(2+p)}/(Kn\mu)^{1/(2+p)}$ in Theorem 41 to obtain $\widetilde{O}\big((Kn)^{(p+2)/(p+4)} \gamma^{-2p/(p+4)}\big)$.

For the parametric case, set $\alpha = \beta = \gamma/Kn$ and $\mu = \sqrt{d \log(Kn/\gamma)/Kn}$ to conclude the bound.

Similarly, in the finite class case, set $\alpha = \beta = 0$ and $\mu = \sqrt{\log|\Pi|/Kn}$. $\square$

Proof of Example 26. Let $\mathcal{F}|_a = \{x \mapsto f(x)_a \mid f \in \mathcal{F}\}$. Then clearly it holds that

$$\log \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}, n) \leq \sum_{a \in [K]} \log \mathcal{N}_{\infty}(\varepsilon, \mathcal{F}|_a, n) \leq K \max_{a \in [K]} \log \mathcal{N}_{\infty}(\varepsilon, \mathcal{F}|_a, n),$$

where have dropped the second "∞" subscript on the right-hand side to denote that this is the covering number for a scalar-valued class. Let $a^{\star}$ be the action that obtains the maximum in this expression. Returning to the integral expression in Theorem 41, we have just shown an upper bound of

$$3e^2 \alpha K n + 12eK \sqrt{\frac{n}{\mu}} \int_{\alpha}^{\beta} \sqrt{\log \mathcal{N}_{\infty}(\varepsilon, \mathcal{F}|_{a^{\star}}, n)} d\varepsilon.$$

For any scalar-value function class $\mathcal{G} \subseteq (\mathcal{X} \to [0,1])$, define

$$\mathcal{R}^{\mathrm{seq}}(\mathcal{G}) = \sup_{\boldsymbol{x}} \mathbb{E}_{\epsilon} \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \epsilon_t g(\boldsymbol{x}_t(\epsilon)).$$

Following the proof of Lemma 9 in Rakhlin et al. (2015), by choosing $\beta = 1$ and $\alpha = 2\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star})/n$, we may upper bound the L_∞ covering number by the sequential Rademacher complexity (via fat-shattering), to obtain

$$6eK\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star}) + 96\sqrt{2}eK\sqrt{\frac{1}{\mu}\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star})} \int_{2\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star})/n}^{1} \frac{1}{\varepsilon}\sqrt{\log(2en/\varepsilon)}d\varepsilon.$$

Using straightforward calculation from the proof of Lemma 9 in Rakhlin et al. (2015), this is upper bounded by

$$O\left(\frac{K}{\sqrt{\mu}}\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star}) \log^{3/2}(n/\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star}))\right).$$

Returning to the regret bound in Theorem 41, we have shown an upper bound of

$$O\left(\frac{K}{\gamma\sqrt{\mu}}\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star}) \log^{3/2}(n/\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star})) + \mu K n\right),$$

where we have used that $\log \mathcal{N}_\infty(1, \mathcal{F}|_{a^\star}, n) = 0$ under the boundedness assumption on $\mathcal{F}$. Setting $\mu \propto (\mathcal{R}^{\mathrm{seq}}(\mathcal{F}|_{a^\star})/(n\gamma))^{2/3}$ yields the result. □

Proof of Example 27. This is an immediate consequence of Example 26 and that Banach spaces for which the martingale type property holds with constant β have sequential Rademacher complexity $O(\sqrt{\beta n})$ (Srebro et al., 2011). □

11.5.6 Additional Minimax Results

Here we briefly state an analogue of Theorem 41 for the hinge loss. Note that this bound leads to the same exponents for n as Theorem 41, but has worse dependence on the margin γ and depends on the scale parameter B explicitly.

Theorem 45 (Contextual bandit chaining bound for hinge loss)**.** *For any fixed constants $\beta > \alpha > 0$, hinge loss parameter $\gamma > 0$, and smoothing parameter $\mu \in (0, 1/K]$ there exists an adversarial contextual bandit strategy $(P_t)_{t\le n}$ with expected regret bounded as*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\ell_t(a_t)\right] \\
&\le \frac{1}{K}\left\{\inf_{f\in\mathcal{F}}\mathbb{E}\left[\sum_{t=1}^{n}\langle\psi^\gamma(f(x_t)),\ell_t\rangle\right] + \frac{1}{\gamma}\sqrt{2K^2B^2n\log\mathcal{N}_{\infty,\infty}(\beta/2,\mathcal{F},n)} + \mu BK^2 n\right. \\
&\qquad\left. + \frac{8B}{\gamma\mu}\log\mathcal{N}_{\infty,\infty}(\beta/2,\mathcal{F},n) + \frac{1}{\gamma}\left(3e\alpha Kn + 24e\sqrt{\frac{Kn}{\mu}}\int_{\alpha}^{\beta}\sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},n)}d\varepsilon\right)\right\},
\end{aligned}$$

where we recall $B = \sup_{f\in\mathcal{F}}\sup_{f\in\mathcal{X}}\|f(x)\|_\infty$.

11.6 Detailed Proofs for Algorithmic Results

11.6.1 Analysis of HINGE-LMC

This section contains the proofs of Theorem 42 and the corresponding corollaries. The proof has many ingredients which we compartmentalize into subsections. First, in Section 11.6.2, we analyze the sampling routine, showing that Langevin Monte Carlo can be used to generate a sample from an approximation of the exponential weights distribution. Then, in Section 11.6.3, we derive the regret bound for the continuous version of exponential weights. Finally, we put the components together together, instantiate all parameters, and compute the final regret and running time in Section 11.6.4. The corollaries are straightforward and proved in Section 11.6.5

To begin, we restate the main theorem, with all the assumptions and the precise parameter settings.

Theorem 46. *Let $\mathcal{F}$ be a set of functions parameterized by a compact convex set $\Theta \subset \mathbb{R}^d$ that* *contains the origin-centered Euclidean ball of radius* 1 *and is contained within a Euclidean ball of radius* R*. Assume that* $f(x;\theta)$ *is convex in* θ *for each* $x \in \mathcal{X}$*, and that* $\sup_{x,\theta} \|f(x;\theta)\|_\infty \le B$*, that* $f(x,a;\theta)$ *is* L*-Lipschitz as a function of* θ *with respect to the* ℓ_2 *norm for each* x, a*. For any* γ*, if we set*

$$\eta = \sqrt{\frac{d\gamma^2 \log(RLnK/\gamma)}{5K^2B^2n}}, \qquad \mu = \sqrt{\frac{1}{K^2n}}, \qquad M = \sqrt{n},$$

in HINGE-LMC*, and further set*

$$u = \frac{1}{n^{3/2}LB_\ell R\eta\sqrt{d}}, \qquad \lambda = \frac{1}{8n^{1/2}R^3}, \qquad \alpha = \frac{R^2}{N},$$
$$N = \tilde{O}\left(R^{18}L^{12}n^6d^{12} + \frac{R^{24}L^{48}d^{12}}{K^{24}}\right), \qquad m = \tilde{O}\left(n^3dR^4L^2B_\ell^2(K\gamma)^{-2}\right),$$

in each call to Projected LMC, then HINGE-LMC *guarantees*

$$\begin{aligned}\sum_{t=1}^{n}\mathbb{E}\ell_t(a_t) &\le \min_{\theta\in\Theta}\sum_{t=1}^{n}\mathbb{E}\langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle + \frac{\sqrt{n}}{\gamma} + \frac{2d}{K\eta}\log(RLnK/\gamma) + \frac{10\eta}{\gamma^2}B^2Kn \\ &\le \min_{\theta\in\Theta}\sum_{t=1}^{n}\mathbb{E}\langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle + \tilde{O}\left(\frac{B}{\gamma}\sqrt{dn}\right).\end{aligned}$$

Moreover, the running time of HINGE-LMC *is* $\tilde{O}\Big(\frac{R^{22}L^{14}d^{14}B_\ell^2n^{10}}{K^2\gamma^2} + \frac{R^{28}L^{50}d^{14}B_\ell^2n^4}{K^{26}\gamma^2}\Big)$*.*

11.6.2 Analysis of the Sampling Routine

In this section, we show how Projected LMC can be used to generate a sample from a distribution that is close to the exponential weights distribution. Define

$$F(\theta) = \eta\sum_{\tau=1}^{t}\langle\tilde{\ell}_\tau, \psi^\gamma(f(x_\tau;\theta))\rangle, \quad P(\theta) \propto \exp(-F(\theta)). \tag{11.11}$$

We are interested in sampling from $P(\theta)$.

Algorithm 14 Smoothed Projected Langevin Monte Carlo for (11.11)

Input: Parameters m, u, λ, N, α.
Set $\tilde{\theta}_0 \leftarrow 0 \in \mathbb{R}^d$
for $k = 1, \ldots, N$ **do**
Sample $z_1, \ldots, z_m \overset{iid}{\sim} \mathcal{N}(0, u^2 I_d)$ and form the function

$$\tilde{F}_k(\theta) = \frac{1}{m}\sum_{i=1}^{m} F(\theta + z_i) + \frac{\lambda}{2}\|\theta\|_2^2.$$

Sample $\xi_k \sim \mathcal{N}(0, I_d)$ and update

$$\tilde{\theta}_k \leftarrow \mathcal{P}_\Theta\left(\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla \tilde{F}_k(\tilde{\theta}_{k-1}) + \sqrt{\alpha}\xi_k\right).$$

end for
Return $\tilde{\theta}_N$.

Let us define the Wasserstein distance. For random variables X, Y with density μ, ν respectively

$$\mathcal{W}_1(\mu, \nu) \triangleq \inf_{\pi \in \Gamma(\mu,\nu)} \int \|X - Y\|_2 d\pi(X, Y) = \sup_{f:\mathrm{Lip}(f)\le 1} \left| \int f(d\mu(X) - d\nu(Y)) \right|.$$

Here $\Gamma(\mu, \nu)$ is the set of couplings between the two densities, that is the set of joint distributions with marginals equal to μ, ν. $\mathrm{Lip}(f)$ is the set of all functions that are 1-Lipschitz with respect to ℓ_2.

Theorem 47. *Let $\Theta \subset \mathbb{R}^d$ be a convex set containing a Euclidean ball of radius $r = 1$ with center 0, and contained within a Euclidean ball of radius R. Let $f : \mathcal{X} \times \Theta \to \mathbb{R}^K_{=0}$ be convex in θ with $f_a(x; \cdot)$ being L-Lipschitz w.r.t. ℓ_2 norm for each $a \in \mathcal{A}$. Assume $\|\tilde{\ell}_\tau\|_1 \le B_\ell$ and define F and P as in (11.11). Let a target accuracy $\tau > 0$ be fixed. Then Algorithm 14 with parameters $m, N, \lambda, u, \alpha \in poly(1/\tau, d, R, \eta, B_\ell, L)$ generates a sample from a distribution $\tilde{P}$ satisfying*

$$\mathcal{W}_1(\tilde{P}, P) \le \tau.$$

Therefore, the algorithm runs in polynomial time.

The precise values for each of the parameters m, N, u, λ, α can be found at the end of the proof, which will lead to a setting of τ in application of the theorem.

Towards the proof, we will introduce the intermediate function $\hat{F}(\theta) = \mathbb{E}_Z F(\theta + Z) + \frac{\lambda}{2}\|\theta\|_2^2$, where Z is a random variable with distribution $\mathcal{N}(0, u^2 I_d)$. This is the randomized smoothing technique studied by Duchi, Bartlett and Wainwright (Duchi et al., 2012). The critical properties of this function are

Proposition 20 (Properties of $\hat{F}$)**.** Under the assumptions of Theorem 47, The function $\hat{F}$ satisfies

1. $F(\theta) \le \hat{F}(\theta) \le F(\theta) + \eta n B_\ell L u\sqrt{d}/\gamma + \frac{\lambda}{2}R^2$.
2. $\hat{F}(\theta)$ is $\eta n B_\ell L/\gamma + \lambda R$-Lipschitz with respect to the ℓ_2 norm.
3. $\hat{F}(\theta)$ is continuously differentiable and its gradient is $\frac{\eta n B_\ell L}{u\gamma} + \lambda$-Lipschitz continuous with respect to the ℓ_2 norm.
4. $\hat{F}(\theta)$ is λ-strongly convex with respect to the ℓ_2 norm.
5. $\mathbb{E}\nabla F(\theta + Z) = \nabla \hat{F}(\theta)$.

Proof. See Duchi et al. (2012, Lemma E.3) for the proof of all claims, except for claim 4, which is an immediate consequence of the ℓ_2 regularization term. □

Using property 1 in Proposition 20 and setting $\varepsilon_1 \triangleq \eta n B_\ell L u\sqrt{d}/\gamma + \lambda R^2$, we know that

$$e^{-\varepsilon_1}\exp(-F(\theta) \le \exp(-\hat{F}(\theta)) \le \exp(-F(\theta)),$$

pointwise. Therefore, defining $\hat{P}$ to be the distribution with density $\hat{p}(\theta) = \exp(-\hat{F}(\theta))/\hat{Z}$, where $\hat{Z} = \int \exp(-\hat{F}(\theta))d\theta$, we have

$$TV(P \,||\, \hat{P}) = \int \frac{e^{-F(\theta)}}{Z}\left|\frac{e^{-\hat{F}(\theta)+F(\theta)}}{\hat{Z}/Z} - 1\right| d\theta \le e^{\varepsilon_1} - 1 \le 2\varepsilon_1,$$

for $\varepsilon_1 \leq 1$. This shows that $\hat{P}$ approximates P well when u and λ are sufficiently small. The next lemma further shows that the $\tilde{F}_k$ functions themselves approximate $\hat{F}$ well.

Lemma 37 (Properties of $\tilde{F}_k$)**.** For any fixed θ, $k \in [N]$, and constant $\varepsilon_2 > 0$,

$$\mathbb{P}\left[\left\|\nabla \hat{F}(\theta) - \nabla \tilde{F}_k(\theta)\right\|_2 \geq \varepsilon_2 + \frac{2}{\sqrt{m}} \cdot \frac{\eta n B_\ell L}{\gamma}\right] \leq \exp\left(\frac{-4\varepsilon_2^2 \gamma^2 m}{(\eta n L B_\ell)^2}\right).$$

Proof of Lemma 37. Let k be fixed. Since $\tilde{F}_k$ are identically distributed for all k we will henceforth abbreviate to $\tilde{F}$.

We proceed using a crude concentration argument. Observe that by Proposition 20, $\mathbb{E}\nabla\tilde{F}(\theta) = \nabla\hat{F}(\theta)$ and moreover $\nabla\tilde{F}(\theta)$ is a sum of m i.i.d., vector-valued random variables (plus the deterministic regularization term).

Via the Chernoff method, for any fixed θ, we have

$$\mathbb{P}\left[\|\nabla\tilde{F}(\theta) - \nabla\hat{F}(\theta)\|_2 \geq t\right] \leq \inf_{\beta>0} \exp(-t\beta)\mathbb{E}\exp(\beta\|\nabla\tilde{F}(\theta) - \nabla\hat{F}(\theta)\|_2)$$

Using the sum structure and symmetrizing:

$$\leq \inf_{\beta>0} \exp(-t\beta)\mathbb{E}_{z_{1:m}} \mathop{\mathbb{E}}_{\epsilon} \exp\left(2\beta\left\|\frac{1}{m}\sum_{i=1}^{m} \epsilon_t \nabla G(\theta + z_i)\right\|_2\right),$$

where $G(\theta) = \eta\sum_{\tau=1}^{t}\langle\tilde{\ell}_\tau, \psi^\gamma(f(x_\tau;\theta))\rangle$. Condition on $z_{1:m}$ and let $W(\epsilon) = \left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i \nabla G(\theta + z_i)\right\|_2$. Then for any i,

$$\begin{aligned}
|W(\epsilon_1, \ldots, \epsilon_i, \ldots, \epsilon_m) - W(\epsilon_1, \ldots, -\epsilon_i, \ldots, \epsilon_m)| &\leq \frac{1}{m}\|\nabla G(\theta + z_i)\|_2 \\
&\leq \frac{\eta}{m}\sum_{\tau=1}^{t}\left\|\tilde{\ell}_\tau\right\|_1 \|\nabla\psi^\gamma(f(x_\tau;\theta + z_i))_a\|_2 \\
&\leq \frac{\eta n B_\ell L}{m\gamma}.
\end{aligned}$$

By the standard bounded differences argument (e.g. (Boucheron et al., 2013)), this implies that $W - \mathbb{E}\,W$ is subgaussian with variance proxy $\sigma^2 = \frac{1}{4m}\left(\frac{\eta n B_\ell L}{\gamma}\right)^2$. Furthermore, the standard application of Jensen's inequality implies that $\mathbb{E}\,W \leq 2\sigma$.

Returning to the upper bound, these facts together imply

$$\mathop{\mathbb{E}}_{\epsilon} \exp\left(2\beta\left\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_t \nabla G(\theta+z_i)\right\|_2\right) \le \exp(2\beta^2\sigma^2+4\beta\sigma).$$

The final bound is therefore,

$$\mathbb{P}\left[\|\nabla\tilde{F}(\theta)-\nabla\hat{F}(\theta)\|_2 \ge t\right] \le \inf_{\beta>0}\exp(-t\beta+2\beta^2\sigma^2+4\beta\sigma).$$

Rebinding $t = t'+4\sigma$ for $t' \ge 0$, we have

$$\mathbb{P}\left[\|\nabla\tilde{F}(\theta)-\nabla\hat{F}(\theta)\|_2 \ge t'+4\sigma\right] \le \inf_{\beta>0}\exp(-t'\beta+2\beta^2\sigma^2) = \exp(-(t')^2/8\sigma^2).$$

□

Now, for the purposes of the proof, suppose we run the Projected LMC algorithm on the function $\hat{F}$, which generates the iterate sequence $\hat{\theta}_0 = 0$

$$\hat{\theta}_k \leftarrow \mathcal{P}_\Theta\left(\hat{\theta}_{k-1} - \frac{\alpha}{2}\nabla\hat{F}(\hat{\theta}_{k-1}) + \sqrt{\alpha}\xi_k\right).$$

Owing to the smoothness of $\hat{F}$, we may apply the analysis of Projected LMC due to Bubeck, Eldan, and Lehec (Bubeck et al., 2018) to bound the total variation distance between the random variable $\hat{\theta}_N$ and the distribution with density proportional to $\exp(-\hat{F}(\theta))$.

Theorem 48 (Bubeck et al. (2018))**.** *Let $\hat{P}$ be the distribution on Θ with density proportional to $\exp(-\hat{F}(\theta))$. For any $\varepsilon > 0$ and with $\alpha = \tilde{\Theta}(R^2/N)$, we have $TV(\hat{\theta}_N, \hat{P}) \le \varepsilon$ with*

$$N \ge \tilde{\Omega}\left(\frac{R^6 \max\{d, R\eta n B_\ell L/\gamma + R^2\lambda, R(\eta n B_\ell L/(u\gamma)+\lambda)\}^{12}}{\varepsilon^{12}}\right).$$

This specializes the result of Bubeck et al. (2018) to our setting, using the Lipschitz and smoothness constants from Proposition 20.

Unfortunately, since we do not have access to $\hat{F}$ in closed form, we cannot run the Projected LMC algorithm on it exactly. Instead, Algorithm 14 runs LMC on the sequence of approximations $\tilde{F}_k$ and generates the iterate sequence $\tilde{\theta}_k$. The last step in the proof is to relate our iterate sequence $\tilde{\theta}_k$ to a hypothetical iterate sequence $\hat{\theta}_k$ formed by running Projected LMC on the function $\hat{F}$.

Lemma 38. Let ε_2 be fixed. Assume the conditions of Theorem 47—in particular that

$$m \geq 16(\eta n L B_\ell/\gamma)^2 \log(4R/\alpha\varepsilon_2)/\varepsilon_2^2, \qquad \alpha \leq 2(\eta n B_\ell L/(u\gamma) + \lambda)^{-1}.$$

Then for any $k \in [N]$ we have $\mathcal{W}_1(\hat{\theta}_k, \tilde{\theta}_k) \leq k\alpha\varepsilon_2$.

Proof of Lemma 38. The proof is by induction, where the base case is obvious, since $\hat{\theta}_0 = \tilde{\theta}_0$. Now, let $\pi^\star_{k-1}$ denote the optimal coupling for $\tilde{\theta}_{k-1}, \hat{\theta}_{k-1}$ and extend this coupling in the obvious way by sampling $z_1, \ldots, z_m$ i.i.d. and by using the same gaussian random variable ξ_k in both LMC updates. Let $\mathcal{E}_k = \{z_{1:m} : \|\nabla \tilde{F}(\tilde{\theta}_{k-1}) - \nabla \hat{F}(\tilde{\theta}_{k-1})\| \leq \varepsilon_2 + \varepsilon'\}$, where $\varepsilon' := \frac{2}{\sqrt{m}} \cdot \frac{\eta n B_\ell L}{\gamma}$; this is the "good" event in which the samples provide a high-quality approximation to the gradient at $\tilde{\theta}_{k-1}$. We then have

$$\begin{aligned}
&\mathcal{W}_1(\hat{\theta}_k, \tilde{\theta}_k) \\
&= \inf_{\pi \in \Gamma(\hat{\theta}_k, \tilde{\theta}_k)} \int \|\hat{\theta}_k - \tilde{\theta}_k\|_2 d\pi \\
&\leq \int \mathbb{E}_{z_{1:m}} \|\mathcal{P}_\Theta(\hat{\theta}_{k-1} - \frac{\alpha}{2}\nabla \hat{F}(\hat{\theta}_{k-1}) - \sqrt{\alpha}\xi_k) - \mathcal{P}_\Theta(\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla \tilde{F}(\tilde{\theta}_{k-1}) - \sqrt{\alpha}\xi_k)\|_2 d\pi^\star_{k-1} \\
&\leq \int \mathbb{E}_{z_{1:m}} \mathbf{1}\{\mathcal{E}_k\} \|\hat{\theta}_{k-1} - \frac{\alpha}{2}\nabla \hat{F}(\hat{\theta}_{k-1}) - (\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla \tilde{F}(\tilde{\theta}_{k-1}))\|_2 d\pi^\star_{k-1} + 2R \int \mathbb{P}[\mathcal{E}^C_{k-1}] d\pi^\star_{k-1} \\
&\leq \int \mathbb{E}_{z_{1:m}} \mathbf{1}\{\mathcal{E}_k\} \|\hat{\theta}_{k-1} - \frac{\alpha}{2}\nabla \hat{F}(\hat{\theta}_{k-1}) - (\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla \tilde{F}(\tilde{\theta}_{k-1}))\|_2 d\pi^\star_{k-1} + 2R \exp\left(\frac{-4\varepsilon_2^2\gamma^2 m}{(\eta n L B_\ell)^2}\right).
\end{aligned}$$

The first inequality introduces the potentially suboptimal coupling $\pi^\star_{k-1}$. In the second inequality we first use that the projection operator is contractive, and we also use that the domain is contained in a Euclidean ball of radius R, providing a coarse upper bound on the second term. For the third inequality, we apply the concentration argument in Lemma 37.

Working just with the first term, using the event in the indicator, we have

$$\int \mathbb{E}_{z_{1:m}} \mathbf{1}\{\mathcal{E}_k\} \|\hat{\theta}_{k-1} - \frac{\alpha}{2}\nabla \hat{F}(\hat{\theta}_{k-1}) - (\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla \tilde{F}(\tilde{\theta}_{k-1}))\|_2 d\pi^\star_{k-1}$$
$$\leq \int \|\hat{\theta}_{k-1} - \frac{\alpha}{2}\nabla \hat{F}(\hat{\theta}_{k-1}) - (\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla \hat{F}(\tilde{\theta}_{k-1}))\|_2 d\pi^\star_{k-1} + \frac{\alpha(\varepsilon_2 + \varepsilon')}{2}.$$

Now, observe that we are performing one step of gradient descent on $\hat{F}$ from two different starting points, $\hat{\theta}_{k-1}$ and $\tilde{\theta}_{k-1}$. Moreover, we know that $\hat{F}$ is smooth and strongly convex, which implies that the gradient descent update is *contractive*. Thus we will be able to upper bound the first term by $\mathcal{W}_1(\hat{\theta}_{k-1}, \tilde{\theta}_{k-1})$, which will lead to the result.

Here is the argument. Consider two arbitrary points $\theta, \theta' \in \Theta$. Let $G : \theta \to \theta - \alpha/2\nabla \hat{F}(\theta)$ be a vector valued function, and observe that the Jacobian is $I - \alpha/2\nabla^2 \hat{F}(\theta)$. By the mean value theorem, there exists θ'' such that

$$\|\theta - \frac{\alpha}{2}\nabla \hat{F}(\theta) - (\theta' - \frac{\alpha}{2}\nabla \hat{F}(\theta'))\|_2 \leq \|(I - \alpha/2\nabla^2 \hat{F}(\theta''))(\theta - \theta')\|_2$$
$$\leq \|I - \alpha/2\nabla^2 \hat{F}(\theta'')\|_\sigma \|\theta - \theta'\|_2.$$

Now, since $\hat{F}$ is λ-strongly convex and $\eta n B_\ell L/u + \lambda$ smooth, we know that all eigenvalues of $\nabla^2 \hat{F}(\theta'')$ are in the interval $[\lambda, \eta n B_\ell L/(u\gamma) + \lambda]$. Therefore, if $\alpha \leq 2(\eta n B_\ell L/(u\gamma) + \lambda)^{-1} \leq 1/\lambda$, the spectral norm term here is at most 1, implying that gradient descent is contractive. Thus, we get

$$\mathcal{W}_1(\hat{\theta}_k, \tilde{\theta}_k) \leq \int \|\hat{\theta}_{k-1} - \tilde{\theta}_{k-1}\|_2 d\pi^\star_{k-1} + \frac{\alpha(\varepsilon_2 + \varepsilon')}{2} + 2R\exp\left(\frac{-4\varepsilon_2^2\gamma^2 m}{(\eta n L B_\ell)^2}\right)$$
$$\leq \mathcal{W}_1(\hat{\theta}_{k-1}, \tilde{\theta}_{k-1}) + \frac{\alpha}{2}\varepsilon_2 + \frac{\alpha}{\sqrt{m}} \cdot \frac{\eta n B_\ell L}{\gamma} + 2R\exp\left(\frac{-4\varepsilon_2^2\gamma^2 m}{(\eta n L B_\ell)^2}\right).$$

The choice of m ensures that the second and third term together are at most $\alpha\varepsilon_2$, from which the result follows. $\square$

Fact 2. For any two distributions μ, ν on Θ, we have

$$\mathcal{W}_1(\mu, \nu) \leq R \cdot TV(\mu, \nu).$$

Proof. We use the coupling characterization of the total variation distance:

$$\mathcal{W}_1(\mu,\nu) = \inf_{\pi} \int \|\theta - \theta'\|_2 d\pi \le \operatorname{diam}(\Theta) \inf_{\pi} \mathbb{P}_\pi[\theta \neq \theta'] \le R \cdot TV(\mu,\nu). \qquad \square$$

Proof of Theorem 47. By the triangle inequality and 2 we have

$$\mathcal{W}_1(\tilde{\theta}_N, P) \le \mathcal{W}_1(\tilde{\theta}_N, \hat{\theta}_N) + R \cdot \left(TV(\hat{\theta}_N, \hat{P}) + TV(\hat{P}, P) \right).$$

The first term here is the Wasserstein distance between our true iterates $\tilde{\theta}_N$ and the idealized iterates from running LMC on $\hat{F}$, which is controlled by Lemma 38. The second is the total variation distance between the idealized iterates and the smoothed density $\hat{P}$, which is controlled in Theorem 48. Finally, the third term is the approximation error between the smoothed density $\hat{P}$ and the true, non-smooth one P. Together, for any choice of $\varepsilon > 0$ and $\varepsilon_2 > 0$ we obtain the bound

$$\mathcal{W}_1(\tilde{\theta}_N, P) \le N\alpha\varepsilon_2 + R\varepsilon + 2R(\eta n B_\ell L u\sqrt{d}/\gamma + \lambda R^2), \tag{11.12}$$

under the requirements

$$\begin{aligned} N &\ge \frac{c_0 R^6 \max\{d, R\eta n B_\ell L/\gamma + R^2\lambda, R(\eta n B_\ell L/(u\gamma) + \lambda)\}^{12}}{\varepsilon^{12}}, \\ m &\ge \frac{16(\eta n L B_\ell/\gamma)^2 \log(4R/\alpha\varepsilon_2)}{\varepsilon_2^2}. \end{aligned} \tag{11.13}$$

There are also two requirements on α, one arising from Theorem 48 and the other from Lemma 38. These are:

$$\alpha \le 2(\eta n B_\ell L/(u\gamma) + \lambda)^{-1}, \quad \text{and} \quad \alpha = c_1 R^2/N, \tag{11.14}$$

for any constant c_1.

Returning to the error bound, if we set

$$u = \frac{\tau}{8R\eta n B_\ell L\sqrt{d}}, \quad \text{and} \quad \lambda = \frac{\tau}{8R^3},$$

the last term in (11.12) is at most $\tau/2$.

We will make the choice $\alpha = c_1 R^2/N$. In this case, the values for u and λ above, combined with the inequality (11.14) give the constraint

$$N \geq 2c_1 R^2 \cdot \left(\frac{8(\eta n L B_\ell)^2 R\sqrt{d}}{\gamma\tau} + \frac{\tau}{8R^2} \right). \tag{11.15}$$

Now for the first term in (11.12), plug in the choice $\alpha = c_1 R^2/N$ and set $\varepsilon_2 = \tau/(4c_1 R^2)$ so that this term is at most $\tau/4$. For the second term, set $\varepsilon = \tau/(4R)$ so that this term is also at most $\tau/4$. With these choices, the requirements on m and N become:

1) $m \geq \dfrac{64c_1^2 R^4(\eta n B_\ell/\gamma)^2 \log(\tau/(16RN))}{\tau^2}$.

2) $N \geq c_0' R^{18} \max\{d, (R\eta n B_\ell L/\gamma)^2\sqrt{d}/\tau\}^{12}/\tau^{12}$.

Note that the first constraint (11.13) clearly implies the second constraint (11.15), and this proves the theorem. □

11.6.3 Continuous Exponential Weights.

The focus of this section is Lemma 39, which analyzes a continuous version of the Hedge/exponential weights algorithms in the full information setting. This lemma appears in various forms in several places, e.g. Cesa-Bianchi and Lugosi (2006). For the setup, consider an online learning problem with a parametric benchmark class $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta\}$ where $f(\cdot;\theta) \in (\mathcal{X} \to \mathbb{R}^K_{=0})$ and further assume that $\Theta \in \mathbb{R}^d$ contains the centered Euclidean ball of radius $r = 1$ and is contained in the Euclidean ball of radius R. Finally, assume that $f(x;\cdot)_a$ is L-Lipschitz with respect to ℓ_2 norm in θ for all $x \in \mathcal{X}$. On each round t an adversary chooses a context $x_t \in \mathcal{X}$ and a loss vector $\ell_t \in \mathbb{R}^K_+$, the learner then choose a distribution $p_t \in \Delta(\mathcal{F})$ and suffers loss:

$$\mathbb{E}_{f\sim p_t}\langle \ell_t, \psi^\gamma(f(x_t))\rangle.$$

The entire loss vector ℓ_t is then revealed to the learner. Here, performance is measured via regret:

$$\mathrm{Reg}_n(n, \mathcal{F}) \triangleq \sum_{t=1}^n \mathbb{E}_{f\sim p_t}\langle \ell_t, \psi^\gamma(f(x_t))\rangle - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \langle \ell_t, \psi^\gamma(f(x_t))\rangle.$$

Our algorithm is a continuous version of exponential weights. Starting with $w_0(f) \triangleq 0$, we perform the updates:

$$p_t(f) = \frac{\exp(-\eta w_t(f))}{\int_{\mathcal{F}} \exp(-\eta w_t(f)) d\lambda(f)}, \quad \text{and} \quad w_{t+1}(f) = w_t(f) + \langle \ell_t, \psi^\gamma(f(x_t))\rangle.$$

Here η is the learning rate and λ is the Lebesgue measure on $\mathcal{F}$ (identifying elements $f \in \mathcal{F}$ with their representatives $\theta \in \mathbb{R}^d$).

With these definitions, the continuous Hedge algorithm enjoys the following guarantee.

Lemma 39. Assume that the losses ℓ_t satisfy $\|\ell_t\|_\infty \leq B_\ell$, $\Theta \subset \mathbb{R}^d$ is contained within the Euclidean ball of radius R, and $f(x; \cdot)_a$ is L-Lipschitz continuous in the third argument with respect to ℓ_2. Let the margin parameter γ be fixed. Then the continuous Hedge algorithm with learning rate $\eta > 0$ enjoys the following regret guarantee:

$$\mathrm{Regret}(n, \mathcal{F}) \leq \inf_{\varepsilon>0}\left\{\frac{nKB_\ell\varepsilon}{\gamma} + \frac{d}{\eta}\log(RL/\varepsilon) + \frac{\eta}{2}\sum_{t=1}^n \mathbb{E}_{f\sim p_t}\langle \ell_t, \psi^\gamma(f(x_t))\rangle^2\right\}.$$

Proof. Following the standard analysis for continuous Hedge (e.g. Lemma 10 in Narayanan and Rakhlin (2017)), we know that the regret to some benchmark distribution $Q \in \Delta(\mathcal{F})$ is

$$\sum_{t=1}^n (\mathbb{E}_{f\sim p_t} - \mathbb{E}_{f\sim Q})(\langle \ell_t, \psi^\gamma(f(x_t))\rangle = \frac{\mathrm{KL}(Q \mid\mid p_0) - \mathrm{KL}(Q \mid\mid p_n)}{\eta} + \frac{1}{\eta}\sum_{t=1}^n \mathrm{KL}(p_{t-1} \mid\mid p_t).$$

For the KL terms, using the standard variational representation, we have

$$\begin{aligned}\mathrm{KL}(p_{t-1} \,||\, p_t) &= \log \mathbb{E}_{f\sim p_{t-1}} \exp\left(-\eta\left\langle \ell_t, \psi^\gamma(f(x_t)) - \mathbb{E}_{f\sim p_{t-1}}\psi^\gamma(f(x_t))\right\rangle\right)\\ &\le \log\left(1 + \frac{\eta^2}{2}\mathbb{E}_{f\sim p_{t-1}}\left\langle \ell_t, \psi^\gamma(f(x_t)) - \mathbb{E}_{f\sim p_{t-1}}\psi^\gamma(f(x_t))\right\rangle^2\right)\\ &\le \frac{\eta^2}{2}\mathbb{E}_{f\sim p_{t-1}}\langle \ell_t, \psi^\gamma(f(x_t))\rangle^2.\end{aligned}$$

Here the first inequality is $e^{-x} \le 1 - x + x^2/2$, using that the term inside the exponential is centered. The second inequality is $\log(1+x) \le x$.

Using non-negativity of KL, we only have to worry about the $\mathrm{KL}(Q \,||\, p_0)$ term. Let $f^\star$ be the minimizer of the cumulative hinge loss. Let $\theta^\star \in \Theta$ be a representative for $f^\star$ and let Q be the uniform distribution on $\mathcal{F}_\varepsilon(\theta^\star, x_{1:n}) \triangleq \{\theta : \max_{t\in[n]} \|f(x_t;\theta) - f(x_t;\theta^\star)\|_\infty \le \varepsilon\}$, then we have that $\mathrm{KL}(Q \,||\, p_0)$ is equal to

$$\int_f q(f)\log(q(f)/p_0(f))d\lambda(f) = \int dQ(f)\cdot \log\frac{\int_{\mathcal{F}} d\lambda(f)}{\int_{\mathcal{F}_\varepsilon} d\lambda(f)} = \log\frac{\text{Vol}(\mathcal{F})}{\text{Vol}(\mathcal{F}_\varepsilon(\theta^\star, x_{1:n}))},$$

where Vol(S) denotes volume under the Lebesgue integral. We know that Vol$(\Theta) \le c_d R^d$ where c_d is the Lebesgue volume of the unit Euclidean ball and R is the radius of the ball containing Θ, and so we must lower bound the volume of $\mathcal{F}_\varepsilon(f^\star, x_{1:n})$. For this step, observe that by the Lipschitz-property of f,

$$\sup_{x\in\mathcal{X}} \|f(x;\theta) - f(x;\theta^\star)\|_\infty \le L\|\theta - \theta^\star\|_2,$$

and hence $\mathcal{F}_\varepsilon(\theta^\star, x_{1:n}) \supset B_2(\theta^\star, \varepsilon/L)$. Thus the volume ratio is

$$\frac{\text{Vol}(\mathcal{F})}{\text{Vol}(\mathcal{F}_\varepsilon(\theta^\star, x_{1:n}))} \le \frac{c_d R^d}{c_d(\varepsilon/L)^d} = (RL/\varepsilon)^d.$$

Finally, using the fact that the hinge surrogate is $1/\gamma$-Lipschitz, we know that

$$\begin{aligned}\sum_{t=1}^n \mathbb{E}_{f\sim Q}\langle \ell_n, \psi^\gamma(f(x_t)) - \psi^\gamma(f^\star(x_t))\rangle &\le nB_\ell \sup_{t\in[n], f\in\text{supp}(Q)} \|\psi^\gamma(f(x_t)) - \psi^\gamma(f^\star(x_t))\|_1\\ &\le \frac{nKB_\ell\varepsilon}{\gamma}.\end{aligned}$$

□

11.6.4 From Full Information to Bandits.

We now combine the results of Section 11.6.2 and Section 11.6.3 to give the final guarantee for HINGE-LMC.

We begin by translating the regret bound in Lemma 39, followed by many steps of approximation. At round t, let P_t denote the Hedge distribution on Θ using the losses $\tilde{\ell}_{1:t-1}$. Let $\tilde{P}_t$ denote the distribution from which $\theta_t \in \Theta$ is sampled in Algorithm 14.

Let $p_t \in \Delta(\mathcal{A})$ denote the induced distributions on actions induced by P_t, i.e. the distribution induced by the process $\theta \sim P_t$, $p_t(a) \propto \psi^\gamma(f(x_t;\theta))$. Likewise, let $\sim p_t \in \Delta(\mathcal{A})$ be the distribution induced by $\theta \sim \tilde{P}_t$, $\tilde{p}_t(a) \propto \psi^\gamma(f(x_t;\theta))$; in this notation $\tilde{p}_t^\mu$ is precisely the distribution from which actions are sampled in Algorithm 12.

Recall that we use μ in the superscript to denote smoothing (e.g. p_t^μ). Let m_t denote the random variable sampled at round t to approximate the importance weight.

We also let $\hat{\ell}_t(a) = \frac{\ell_t(a)}{\tilde{p}_t^\mu(a)}\mathbf{1}\{a_t = a\}$ denote estimated losses under the true importance weights, which are not explicitly used by Algorithm 12 but are used in the analysis.

Let $\mathbf{1}_a \in \mathbb{R}^K$ be the vector with 1 at coordinate a and 0 at all other coordinates.

Proof of Theorem 46. The thrust of this proof is to show that the full information bound in Lemma 39 does not degrade significantly under importance weighting, nor does it degrade under the approximate LMC implementation of continuous exponential weights.
Variance control Controlling the variance term in Lemma 39 requires an application of Lemma 32. After taking conditional expectations, the variance term is

$$\sum_{t=1}^{n} \mathbb{E}_{\theta\sim P_t}\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\mathbb{E}_{m_t}\langle\tilde{\ell}_t, \psi^\gamma(f(x_t;\theta))\rangle^2 = \sum_{t=1}^{n} \mathbb{E}_{s\sim P_t}\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\mathbb{E}_{m_t} m_t^2\langle\ell_t(a_t)\mathbf{1}_{a_t}, s\rangle^2.$$

Here we are identifying s with $\psi^\gamma(f(x_t;\theta))$ and marginalizing out θ in the outermost expectation. Note that this is the same definition of s as in Lemma 32.

First let us handle the m_t random variable. Note that conditional on everything up to round t and a_t, the variable m_t is distributed according to a geometric distribution with mean $\tilde{p}_t^\mu(a_t)$, truncated at M. It is straightforward (cf. Neu and Bartók (2013)) to show that m_t is stochastically dominated by a geometric random variable with mean $\frac{1}{\tilde{p}_t^\mu(a_t)}$ and hence the second moment of this random variable is at most $\frac{2}{\tilde{p}_t^\mu(a_t)^2}$. Thus, we are left with

$$\begin{aligned}
&\leq 2\sum_{t=1}^{n} \mathbb{E}_{s\sim P_t}\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\frac{1}{\tilde{p}_t^\mu(a_t)^2}\langle \ell_t(a_t)\mathbf{1}_{a_t}, s\rangle^2 \\
&= 2\sum_{t=1}^{n} \mathbb{E}_{s\sim P_t}\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\langle \hat{\ell}_t, s\rangle^2 \\
&\leq 2\sum_{t=1}^{n} (\mathbb{E}_{s\sim\tilde{P}_t} - \mathbb{E}_{s\sim P_t})\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\langle \hat{\ell}_t, s\rangle^2 + \mathbb{E}_{s\sim\tilde{P}_t}\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\langle \hat{\ell}_t, s\rangle^2.
\end{aligned}$$

We can apply Lemma 32 on the second term, since the only condition for the lemma is that the action distribution is induced from the distribution in the outer expectation. It follows that this term is bounded as

$$\sum_{t=1}^{n} \mathbb{E}_{s\sim\tilde{P}_t}\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\langle \hat{\ell}_t, s\rangle^2 \leq nK^2(1+B/\gamma)^2.$$

For the first term, evaluating the inner expectation, using the fact that $\tilde{p}_t^\mu(a) \geq \mu$ and applying the Lipschitz properties of $\psi^\gamma(\cdot)$, $f(x;\cdot)$ (in particular that $f(x;\cdot)$ is L-Lipschitz with respect to ℓ_2 and that the Wasserstein distance we work with is defined relative to ℓ_2) we have

$$\begin{aligned}
(\mathbb{E}_{s\sim\tilde{P}_t} - \mathbb{E}_{s\sim P_t})\mathbb{E}_{a_t\sim\tilde{p}_t^\mu}\langle \hat{\ell}_t, s\rangle^2 &= \sum_a (\mathbb{E}_{\theta\sim\tilde{P}_t} - \mathbb{E}_{\theta\sim P_t})\frac{\ell_t^2(a)}{\tilde{p}_t^\mu(a)}\psi^\gamma(f(x_t;\theta)_a)^2 \\
&\leq 2\frac{(1+B/\gamma)KL}{\gamma\mu}\sup_{g,\|g\|_{\text{Lip}}\leq 1}\left|\int g(dP_t - d\tilde{P}_t)\right| \\
&= 2\frac{(1+B/\gamma)KL}{\gamma\mu}\mathcal{W}_1(P_t,\tilde{P}_t).
\end{aligned}$$

Finally, using the Wasserstein guarantee $\mathcal{W}_1(P_t, \tilde{P}_t) \le \tau$ from Theorem 47, we conclude that the cumulative variance term is upper bounded as

$$\sum_{t=1}^{n} \mathbb{E}\langle \tilde{\ell}_t, \psi^\gamma(f(x_t;\theta))\rangle^2 \le \frac{4(1+B/\gamma)KnL\tau}{\gamma\mu} + 2(1+B/\gamma)^2K^2n.$$

Bounding regret We first relate the cumulative loss under Algorithm 12 to the cumulative loss of continuous exponential weights. Observe that

$$\begin{aligned}\sum_{t=1}^{n} \langle \ell_t, \tilde{p}_t^\mu\rangle &\le \mu Kn + \sum_{t=1}^{n}\langle \ell_t, \tilde{p}_t\rangle \\ &\le \mu Kn + \frac{1}{K}\sum_{t=1}^{n}\mathbb{E}_{\theta\sim\tilde{P}_t}\langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle \\ &\le \mu Kn + \frac{nL\tau}{\gamma} + \frac{1}{K}\sum_{t=1}^{n}\mathbb{E}_{\theta\sim P_t}\langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle.\end{aligned}$$

This first inequality is a straightforward consequence of smoothing, while the second is a direct application of Lemma 31.

The third inequality is based on the fact that $\langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle$ is KL/γ-Lipschitz in θ with respect to ℓ_2 norm under our assumptions. This step also uses the Wasserstein guarantee in Theorem 47 which produces the approximation factor τ.

Following the analysis in Neu and Bartók (2013) and using the boundedness of ψ^γ, the bias introduced due to using geometric resampling with truncation at M instead of exact inverse propensity scores is

$$\sum_{t=1}^{n}\mathbb{E}_{\theta\sim P_t}\langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle \le \mathbb{E}_{a_{1:n},m_{1:n}}\sum_{t=1}^{n}\mathbb{E}_{\theta\sim P_t}\langle \tilde{\ell}_t, \psi^\gamma(f(x_t;\theta))\rangle + \frac{n(1+B/\gamma)}{eM}.$$

For the remaining term, we apply Lemma 39 with $\varepsilon = \gamma/(nKM)$, since M is an upper bound on the norm $\left\|\tilde{\ell}_t\right\|_1$ of the losses to the full information algorithm.

$$\begin{aligned}
&\mathbb{E}_{a_{1:n},m_{1:n}} \sum_{t=1}^{n} \mathbb{E}_{\theta\sim P_t} \langle \tilde{\ell}_t, \psi^\gamma(f(x_t;\theta))\rangle \\
&\leq \mathbb{E} \inf_{\theta\in\Theta} \sum_{t=1}^{n} \langle \tilde{\ell}_t, \psi^\gamma(f(x_t,\theta))\rangle + 1 + \frac{d}{\eta}\log(RLnKM/\gamma) + \eta\left(\frac{(1+B/\gamma)KnL\tau}{\gamma\mu} + (1+B/\gamma)^2K^2n\right).
\end{aligned}$$

The first term here is the benchmark we want to compare to, since $\mathbb{E}\inf(\cdot) \leq \inf\mathbb{E}[\cdot]$ and so the regret contains several terms:

$$\begin{aligned}
\mu Kn + \frac{nL\tau}{\gamma} + \frac{n(1+B/\gamma)}{eMK} + \frac{1}{K} + \frac{d}{K\eta}\log(RLnKM/\gamma) \\
+ \frac{\eta}{2K}\left(\frac{2(1+B/\gamma)KnL\tau}{\gamma\mu} + 4(1+B/\gamma)^2K^2n\right),
\end{aligned}$$

which simplifies to an upper bound of

$$\leq \mu Kn + \frac{nL\tau}{\gamma} + \frac{n(1+B/\gamma)}{eMK} + \frac{1}{K} + \frac{d}{K\eta}\log(RLnKM/\gamma) + \frac{2\eta}{K\gamma^2}\left(\frac{BKnL\tau}{\mu} + 4B^2K^2n\right).$$

Here we have used the assumption $B/\gamma \geq 1$. We will simplify the expression to obtain an $\tilde{O}(\sqrt{dKn})$-type bound, first set $\mu = 1/(K\sqrt{n})$, $M = \sqrt{n}$ and $\tau = \sqrt{1/(nL^2)}$. This gives

$$\begin{aligned}
&2\sqrt{n} + \frac{2B}{\gamma}\sqrt{n} + \frac{2d}{K\eta}\log(RLnK/\gamma) + \frac{2\eta}{K\gamma^2}\left(BK^2n + 4B^2K^2n\right) \\
&\leq O(B\sqrt{n}/\gamma) + \frac{2d}{K\eta}\log(RLnK/\gamma) + \frac{10\eta}{\gamma^2}B^2Kn.
\end{aligned}$$

Finally set $\eta = \sqrt{\frac{d\gamma^2\log(RLnK/\gamma)}{5K^2B^2n}}$ to get

$$O(\sqrt{n}/\gamma) + O\left(\frac{B}{\gamma}\sqrt{dn\log(RLnK/\gamma)}\right) = \tilde{O}(B\sqrt{dn}/\gamma).$$

This concludes the proof of the regret bound.

Running time calculation. At each round make $M+1$ calls to the LMC sampling routine for a total of $O(n^{3/2})$ calls across all rounds. We now bound the running time for a single call.

We always use parameter $\tau = \sqrt{1/(nL^2)}$ and we know $\|\tilde{\ell}\|_1 \le 1/\mu = K\sqrt{n}$ and $\eta = \tilde{O}(\sqrt{\frac{d}{K^2 n}})$. Plugging into the parameter choices at the end of the proof of Theorem 47, we must sample

$$m = \tilde{O}(n^3 d R^4 L^2 B_\ell^2/(K\gamma)^2)$$

samples from a gaussian distribution on each iteration, and the number of iterations to generate a single sample is:

$$N = \tilde{O}\left(R^{18}L^{12}n^6 d^{12} + \frac{R^{24}L^{48}d^{12}}{K^{24}}\right).$$

Therefore, the total running time across all rounds is

$$\tilde{O}\left(\frac{R^{22}L^{14}d^{14}B_\ell^2 n^{10}}{K^2\gamma^2} + \frac{R^{28}L^{50}d^{14}B_\ell^2 n^4}{K^{26}\gamma^2}\right).$$

□

11.6.5 Proofs for Corollaries

Corollary 13 is an immediate consequence of Theorem 42. For Corollary 14, we apply Lemma 33, since $\theta^\star \in \Theta$ satsifies the conditions of the lemma pointwise. Thus

$$K^{-1}\mathbb{E}\langle \ell_t, \psi^\gamma(f(x_t;\theta^\star))\rangle = K^{-1}\mathbb{E}[\langle \bar{\ell}_t, \psi^\gamma(f(x_t;\theta^\star))\rangle \mid x_t] = \mathbb{E}[\min_a \bar{\ell}_t(a)|x_t].$$

Therefore, letting $a_t^\star$ denote the optimal action minimizing $\bar{\ell}_t$, we obtain the expected regret bound

$$\sum_{t=1}^{n} \mathbb{E}[\langle \bar{\ell}_t, a_t - a_t^\star\rangle] \le \tilde{O}((B/\gamma)\sqrt{dn}).$$

11.6.6 Analysis of SMOOTHFTL

Recall we are in the stochastic setting, and let $\mathcal{D}$ denote the distribution over $(\mathcal{X}, \mathbb{R}_+^K)$ generating the data.

The bulk of the analysis is the following uniform convergence lemma, which is based on chaining for the function class $\mathcal{F}$. Recall that $\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F})$ is the L_∞/ℓ_∞ covering number from Definition 17.

Lemma 40. Fix a predictor $\hat{f}$ and let $\{x_i, a_i, \ell_i(a_i)\}_{i=1}^n$ be a dataset of n samples, Suppose that (x_i, ℓ_i) are drawn i.i.d. from some distribution $\mathcal{D}$ and a_i is sampled from $p_i = (1 - K\mu)\pi_{\text{hinge}}\hat{f}(x_i) + \mu$ for some fixed predictor $\hat{f}$. Define $\hat{R}_n^\psi(f) = \frac{1}{n}\sum_{i=1}^n \langle \hat{\ell}_i, \psi^\gamma(f(x_i))\rangle$, where $\hat{\ell}_i$ is the importance-weighted loss. Then it holds that:

$$\mathbb{E}\sup_{f\in\mathcal{F}} |R^\psi(f) - \hat{R}_n^\psi(f)\rangle|$$
$$\leq \frac{1}{\gamma}\inf_{\beta\geq 0}\left\{2K\beta + 12\int_\beta^2 \left(\sqrt{\frac{2K}{n\mu}\log(n\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},n))} + \frac{3\log(n\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},n))}{n\mu}\right)d\varepsilon\right\}.$$

Proof of Lemma 40. Note that since the data collection policy $\hat{f}$ is fixed, and since we are in the stochastic setting with $(x_i, \ell_i) \sim \mathcal{D}$, the samples $\{x_i, a_i, \ell_i(a_i)\}_{i=1}^n$ are i.i.d. Consequently, we can apply the standard symmetrization upper bound for uniform convergence. Beginning with

$$\mathbb{E}_{x_{1:n},a_{1:n},\ell_{1:n}}\sup_{f\in\mathcal{F}}\Big[R^\psi(f) - \hat{R}_n^\psi(f)\Big],$$

we introduce a second "ghost" dataset of samples $\tau = n+1, \ldots, 2n$ via Jensen's inequality.

$$\leq \mathbb{E}_{x_{1:2n},a_{1:2n},\ell_{1:2n}}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{\tau=n+1}^{2n}\langle\hat{\ell}_\tau, \psi^\gamma(f(x_\tau))\rangle - \frac{1}{n}\sum_{\tau=1}^{n}\langle\hat{\ell}_\tau, \psi^\gamma(f(x_\tau))\rangle.$$

Introducing Rademacher random variables and splitting the supremum:

$$\leq 2\mathbb{E}_{x_{1:n},a_{1:n},\ell_{1:n},\epsilon_{1:n}}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{\tau=1}^{n}\epsilon_\tau\langle\hat{\ell}_\tau, \psi^\gamma(f(x_\tau))\rangle.$$

Now condition on $x_{1:n}$ and define a sequence $\beta_i = 2^{1-i}$ for $i \in \{0, 1, 2, \ldots, N\}$, where N is such that $\beta_{N+1} \geq \beta \geq \beta_{N+2}$ for the value of β in the lemma statement. For each β_i let V_i be a (classical) L_∞/ℓ_∞ cover for f at scale β_i on $x_{1:n}$, that is

$$\forall f \in \mathcal{F}, \forall i, \exists v \in V_i \text{ s.t. } \max_{t \in [n]} \|f(x_t) - v_t\|_\infty \leq \beta_i.$$

We can always ensure $|V_i| \leq \mathcal{N}_{\infty,\infty}(\beta_i, \mathcal{F}, n)$ and since $\|f(x)\|_\infty \leq 1$, we know that $\mathcal{N}_{\infty,\infty}(\beta_0, \mathcal{F}, n) \leq 1$. Now, letting $v^{(i)}(f)$ denote the covering element for f at scale β_i, we have

$$\begin{aligned}
&\mathbb{E}_{a_{1:n},\ell_{1:n},\epsilon_{1:n}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(f(x_\tau)) \rangle \\
&\leq \mathbb{E}_{a_{1:n},\ell_{1:n},\epsilon_{1:n}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(f(x_\tau)) - \psi^\gamma(v_\tau^{(N)}(f)) \rangle \\
&\quad + \sum_{i=1}^{N} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(i)}(f)) - \psi^\gamma(v_\tau^{(i-1)}(f)) \rangle \\
&\quad + \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(0)}(f)) \rangle.
\end{aligned}$$

Since $|V_0| \leq 1$, the expected value of the third term is zero. The remaining work is to bound the first and second terms.

For the first term note that by Hölder's inequality, for any $f \in \mathcal{F}$,

$$\begin{aligned}
\frac{1}{n} \sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(f(x_\tau)) - \psi^\gamma(v_\tau^{(N)}(f)) \rangle &\leq \frac{1}{n} \sum_{\tau=1}^{n} \|\hat{\ell}_\tau\|_1 \|\psi^\gamma(f(x_\tau)) - \psi^\gamma(v_\tau^{(N)}(f))\|_\infty \\
&\leq \frac{\beta_N}{\gamma} \frac{1}{n} \sum_{\tau=1}^{n} \|\hat{\ell}_\tau\|_1,
\end{aligned}$$

since ψ^γ is $1/\gamma$-Lipschitz. Thus for the first term, we have

$$\mathbb{E}_{a_{1:n},\ell_{1:n},\epsilon_{1:n}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(f(x_\tau)) - \psi^\gamma(v_\tau^{(N)}(x_\tau)) \rangle \leq \frac{\beta_N}{\gamma} \mathbb{E}_{a_{1:n},\ell_{1:n}} \frac{1}{n} \sum_{\tau=1}^{n} \|\hat{\ell}_\tau\|_1 \leq \frac{\beta_N K}{\gamma}.$$

Note that there is no dependence on the smoothing parameter μ here.

For the second term, let us denote the ith term in the summation by

$$\mathbb{E}_{a_{1:n},\ell_{1:n},\epsilon_{1:n}} \underbrace{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(i)}(f)) - \psi^\gamma(v_\tau^{(i-1)}(f))\rangle}_{\triangleq\, \mathcal{E}_i}.$$

We control $\mathcal{E}_i$ using Bernstein's inequality and a union bound. First, note that the individual elements in the sum satisfy the deterministic bound

$$|\epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(i)}(f)) - \psi^\gamma(v_\tau^{(i-1)}(f))\rangle| \le \frac{3\beta_i}{\mu\gamma}, \tag{11.16}$$

and the variance bound,

$$\begin{aligned}\mathbb{E}\langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(i)}(f)) - \psi^\gamma(v_\tau^{(i-1)}(f))\rangle^2 &\le \sum_a \mathbb{E}_{a_\tau} \frac{\mathbf{1}\{a_\tau = a\}}{p_\tau(a)^2}(\psi^\gamma(v_\tau^{(i)}(f)_a) - \psi^\gamma(v_\tau^{(i-1)}(f)_a))^2 \\ &\le \sum_a \frac{1}{\mu}(3\beta_i/\gamma)^2 = \frac{9\beta_i^2 K}{\mu\gamma^2}. \end{aligned} \tag{11.17}$$

Here we are using that $v^{(i)}(f)$ and $v^{(i-1)}(f)$ are the covering elements for f, Lipschitzness of ψ^γ, and the definition of the importance weighted loss $\hat{\ell}_\tau$.

Using the bounds (11.16) and (11.17), Bernstein's inequality (e.g. Boucheron et al. (2013), Theorem 2.9) implies that for any $\delta \in (0,1)$,

$$\frac{1}{n}\sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(i)}(f)) - \psi^\gamma(v_\tau^{(i-1)}(f))\rangle \le 6\sqrt{\frac{\beta_i^2 K}{n\mu\gamma^2}\log(1/\delta)} + \frac{6\beta_i}{n\mu\gamma}\log(1/\delta),$$

with probability at least $1-\delta$. The important point here is that $1/(n\mu)$ appears in the square root, as opposed to $1/(n\mu^2)$. Via a union bound, we know that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_f \frac{1}{n}\sum_{\tau=1}^{n} \epsilon_\tau \langle \hat{\ell}_\tau, \psi^\gamma(v_\tau^{(i)}(f)) - \psi^\gamma(v_\tau^{(i-1)}(f))\rangle \\ &\le 6\sqrt{\frac{\beta_i^2 K}{n\mu\gamma^2}\log(|V_i||V_{i-1}|/\delta)} + \frac{6\beta_i}{n\mu\gamma}\log(|V_i||V_{i-1}|/\delta) \\ &\le \frac{6\beta_i}{\gamma}\left(\sqrt{\frac{2K}{n\mu}\log(|V_i|/\delta)} + \frac{2\log(|V_i|/\delta)}{n\mu}\right),\end{aligned}$$

since $|V_{i-1}| \le |V_i|$. Now, recalling the shorthand definition $\mathcal{E}_i$

$$\begin{aligned}
\mathbb{E}_{a_{1:n},\ell_{1:n},\epsilon_{1:n}}\mathcal{E}_i &\le \inf_{\zeta} \mathbb{E}\mathbf{1}\{\mathcal{E}_i \le \zeta\}\cdot\zeta + \mathbb{E}\mathbf{1}\{\mathcal{E}_i > \zeta\}\cdot\frac{3\beta_i}{\mu\gamma} \\
&\le \inf_{\delta\in(0,1)} \frac{6\beta_i}{\gamma}\left(\sqrt{\frac{2K}{n\mu}\log(|V_i|/\delta)} + \frac{2\log(|V_i|/\delta)}{n\mu}\right) + \frac{3\beta_i\delta}{\mu\gamma}.
\end{aligned}$$

Choosing $\delta = 1/n$:

$$\le \frac{6\beta_i}{\gamma}\left(\sqrt{\frac{2K}{n\mu}\log(n|V_i|)} + \frac{3\log(n|V_i|)}{n\mu}\right).$$

Thus, the second term in the chaining decomposition is

$$\begin{aligned}
&\frac{6}{\gamma}\sum_{i=1}^{N}\beta_i\left(\sqrt{\frac{2K}{n\mu}\log(n|V_i|)} + \frac{3\log(n|V_i|)}{n\mu}\right) \\
&= \frac{12}{\gamma}\sum_{i=1}^{N}(\beta_i - \beta_{i+1})\left(\sqrt{\frac{2K}{n\mu}\log(n|V_i|)} + \frac{3\log(n|V_i|)}{n\mu}\right) \\
&\le \frac{12}{\gamma}\int_{\beta_{N+1}}^{\beta_0}\left(\sqrt{\frac{2K}{n\mu}\log(n\mathcal{N}_{\infty,\infty}(\beta,\mathcal{F}))} + \frac{3\log(n\mathcal{N}_{\infty,\infty}(\beta,\mathcal{F}))}{n\mu}\right)d\beta.
\end{aligned}$$

This concludes the uniform deviation statement. Exactly the same argument applies to the other tail, so the bound holds on the absolute value. □

Proof of Theorem 43. Let us denote the right hand side of Lemma 40, when the dataset is size n, as Δ_n. Define,

$$f^\star = \arg\min_{f\in\mathcal{F}} \mathbb{E}\langle \ell, \psi^\gamma(f(x))\rangle,$$

Since the m^{th} epoch proceeds for $n_m \triangleq 2^m$ rounds, and the predictor that we use in the m^{th} epoch is the ERM on all of the data from the $(m-1)^{\text{st}}$ epoch, the expected cumulative hinge regret for the m^{th} epoch is $2^m \cdot \left(\mathbb{E}R^\psi(\hat{f}_{m-1}) - R^\psi(f^\star)\right)$. Using the optimality guarantee for

ERM, this is at most

$$
\begin{aligned}
& 2^m \cdot \left(\mathbb{E} R^\psi(\hat{f}_{m-1}) - \frac{1}{n_{m-1}} \sum_{\tau=n_{m-1}}^{2n_{m-1}-1} \langle \hat{\ell}_\tau, \psi^\gamma(\hat{f}_{m-1}(x_\tau)) \rangle \right) \\
& \quad + 2^m \left(\frac{1}{n_{m-1}} \sum_{\tau=n_{m-1}}^{2n_{m-1}-1} \langle \hat{\ell}_\tau, \psi^\gamma(f^\star(x_\tau)) - R^\psi(f^\star) \right) \\
& \leq 2^{m+1} \mathbb{E} \sup_f \left| R^\psi(f) - \hat{R}^\psi_{n_{m-1}}(f) \right| .
\end{aligned}
$$

Using the guarantee from Lemma 40:

$$
\leq 2^{m+1} \Delta_{n_{m-1}}. \tag{11.18}
$$

Summing this bound over all rounds, the cumulative expected regret after the zero-th epoch is $\sum_{m=1}^{\log_2(n)} 2^{m+1} \Delta_{n_{m-1}}$. The zero-th epoch contributes $1/\gamma$ to the regret, which will be lower order. This gives the following upper bound on the cumulative expected hinge loss regret.

$$
\mathrm{Reg}_n(n, \mathcal{F}) \leq \sum_{m=1}^{\log_2(n)} 2^{m+1} \Delta_{n_{m-1}} \leq \frac{4}{\gamma} \sum_{m=1}^{\log_2(n)} C_m,
$$

where C_m is defined by

$$
\inf_{\beta > 0} \left\{ n_m K \beta + 12 \cdot 2^{m-1} \cdot \int_\beta^{2B} \left(\sqrt{\frac{2K}{n_{m-1}\mu} \log(n_{m-1} \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}))} + \frac{3 \log(n_{m-1} \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}))}{n_{m-1}\mu} \right) d\varepsilon \right\}.
$$

From this definition we have an immediate upper bound of

$$
\begin{aligned}
& \mathrm{Reg}_n(n, \mathcal{F}) \\
& \leq \frac{4}{\gamma} \inf_{\beta > 0} \left\{ K n \beta + 12 \log_2(n) \cdot \int_\beta^{2B} \left(\sqrt{\frac{2Kn}{\mu} \log(n \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}))} + \frac{3 \log(n \mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F}))}{\mu} \right) d\varepsilon \right\} \\
& =: C.
\end{aligned}
$$

Let $z_t = \hat{f}_{m-1}(x_t)$ for each time t in epoch m. We have just shown

$$
\sum_{t=1}^n \mathbb{E} \langle \ell_t, \psi^\gamma(z_t) \rangle \leq n \cdot \mathbb{E} \langle \ell, \psi^\gamma(f^\star(x)) \rangle + C.
$$

Using Lemma 31, this implies

$$\sum_{t=1}^{n} \mathbb{E}\big\langle \ell_t, \pi_{\text{hinge}}(z_t)\big\rangle \leq \frac{n}{K} \cdot \mathbb{E}\langle \ell, \psi^{\gamma}(f^{\star}(x))\rangle + \frac{C}{K}.$$

Finally since $p_t = (1 - K\mu)\pi_{\text{hinge}}(z_t) + \mu$ and $\|\ell_t\|_\infty \leq 1$, this implies the bound

$$\sum_{t=1}^{n} \mathbb{E}\,\ell_t(a_t) \leq \frac{n}{K} \cdot \mathbb{E}\langle \ell, \psi^{\gamma}(f^{\star}(x))\rangle + \underbrace{\frac{C}{K} + \mu K n}_{=:\,C'}.$$

We proceed to bound the final regret C' under the specific covering number behavior assumed in the theorem statement. Assume that $\log(\mathcal{N}_{\infty,\infty}(\varepsilon, \mathcal{F})) \leq \varepsilon^{-p}$ for some $p > 2$. Omitting the $\log(n)$ additive terms, which will contribute $O(B\gamma^{-1}\sqrt{Kn\log(n)/\mu} + B\gamma^{-1}\log(n)/\mu)$ to the overall regret, the bound is now

$$\mu K n + \frac{1}{\gamma K}\left(\inf_{\beta>0} 4Kn\beta + 12\log_2(n) \cdot \int_{\beta}^{2} \sqrt{\frac{2Kn}{\mu\varepsilon^{p}}}d\varepsilon + 36\log_2(n) \cdot \int_{\beta}^{2} \frac{1}{\mu\varepsilon^{p}}d\varepsilon\right).$$

Choosing $\beta = (Kn\mu)^{-1/p}$, this bound becomes

$$O\left(\mu K n + \frac{1}{\gamma K}\log(n)(Kn)^{1-1/p}\mu^{-1/p}\right).$$

Finally, we choose $\mu = \gamma^{-\frac{p}{p+1}} n^{-\frac{1}{p+1}} K^{-1}$, leading to a final bound of

$$O\left((n/\gamma)^{\frac{p}{p+1}}\right).$$

□

11.6.7 SmoothFTL for Lipschitz CB

Here we analyze SmoothFTL in a stochastic Lipschitz contextual bandit setting. To describe the setting, let $\mathcal{X}$ be a metric space endowed with metric ρ and with covering dimension p. This latter fact means that for each $0 < \varepsilon \leq 1$, $\mathcal{X}$ can be covered using at most $C_{\mathcal{X}}\varepsilon^{-p}$ balls of radius ε. Let $\mathcal{A}$ be a finite set of K actions. In this section, we define the benchmark class $\mathcal{G} \subset (\mathcal{X} \to \Delta(\mathcal{A}))$ to be the set of 1-Lipschitz functions, meaning that $\|g(x) - g(x')\|_1 \leq \rho(x, x')$ for all g, x, x' (The choice of ℓ_1 norm is natural since we are operating over the simplex).

We focus on the stochastic setting where there is a distribution $\mathcal{D}$ over $\mathcal{X} \times [0,1]^K$. At each round $(x_t, \ell_t) \sim \mathcal{D}$ is drawn and x_t is presented to the learner. The learner chooses a distribution $p_t \in \Delta(\mathcal{A})$, samples an action $a_t \in \mathcal{A}$ from p_t, and observes the loss $\ell_t(a_t)$. We measure regret via

$$\mathrm{Reg}_n(n, \mathcal{G}) = \sum_{t=1}^{n} \mathbb{E}\ell_t(a_t) - \inf_{g \in \mathcal{G}} n\mathbb{E}\langle g(x), \ell\rangle.$$

In this setting, SMOOTHFTL takes the following form. After the m^{th} epoch, we choose a function $\hat{g}_m$ by solving the empirical risk minimization (ERM) problem

$$\hat{g}_m = \arg\min_{g \in \mathcal{G}} \sum_{\tau = n_m 1}^{2n_m - 1} \langle \hat{\ell}_\tau, g(x_\tau)\rangle,$$

where $\hat{\ell}_\tau$ is the importance weighted loss. Then, we use $\hat{g}_m$ for all the rounds in the subsequent epoch, which means that after observing x_t, we set $p_t(a) = (1 - K\mu)\hat{g}_m(x_t, a) + \mu$. We sample $a_t \sim p_t$, observe $\ell_t(a_t)$ and use the standard importance weighting scheme:

$$\hat{\ell}_t(a) = \frac{\ell_t(a_t)\mathbf{1}\{a = a_t\}}{p_t(a)}.$$

For this algorithm, we have the following theorem.

Theorem 49. SMOOTHFTL *in the Lipschitz CB setting enjoys a regret of* $\tilde{O}((Kn)^{\frac{p}{p+1}})$.

This theorem improves upon the recent result of Cesa-Bianchi et al. (2017), who obtain $\tilde{O}(n^{\frac{p+1}{p+2}})$ in this setting.

Proof. We are in a position to apply Lemma 40. The main difference is that there is no margin parameter, since our functions are 1-Lipschitz, instead of $1/\gamma$-Lipschitz after applying the surrogate loss. The ℓ_∞-metric entropy at scale ε is $C_{\mathcal{X}}\varepsilon^{-p}$ up to polynomial factors in K and logarithmic factors, and so in the m^{th} epoch the ERM has sub-optimality (see (11.18)) at most

$$\tilde{O}\left(\inf_{\beta} K\beta + \int_{\beta}^{1}\sqrt{\frac{K\beta^{-p}}{n_{m-1}\mu}} + \frac{\beta^{-p}}{n_{m-1}\mu}\right),$$

where $\tilde{O}$ hides dependence on $C_{\mathcal{X}}$. Following the argument in the proof of Theorem 43, the overall regret is then

$$\mathrm{Reg}_n(n,\mathcal{G}) = \tilde{O}\left(\mu K n + \inf_{\beta} nK\beta + \int_{\beta}^{1}\sqrt{\frac{nK\beta^{-p}}{\mu}} + \frac{\beta^{-p}}{\mu}\right).$$

Set $\beta = (nK\mu)^{-1/p}$ and then $\mu = (nK)^{\frac{-1}{p+1}}$ now to obtain the result. $\square$

In principle our technique can be further extended to the setting where the action space is also a general metric space, and the losses are Lipschitz, which is the more general setting addressed by Cesa-Bianchi et al. (2017). If the action space has covering dimension $p_{\mathcal{A}}$ then we discretize the action space to resolution ϵ, set $K = \epsilon^{-p_{\mathcal{A}}}$ in the above argument, and balance ϵ with an additional $n\epsilon$ factor that we pay for discretization. This is the approach used in Cesa-Bianchi et al. (2017) to obtain $n^{\frac{p+p_{\mathcal{A}}+1}{p+p_{\mathcal{A}}+2}}$. Unfortunately, our argument above obtains a somewhat poor dependence on K ($K^{\frac{p}{p+1}}$ as opposed to $K^{\frac{1}{p+1}}$, which is more natural). Consequently, the argument produces a bound of $\tilde{O}(n^{\frac{p+pp_{\mathcal{A}}}{p+pp_{\mathcal{A}}+1}})$ which only improves on Cesa-Bianchi et al. (2017) when $p_{\mathcal{A}} \leq 1/(p-1)$.

11.7 Chapter Notes

This chapter is based on Foster et al. (2018b).

Detailed Discussion of Related Work Contextual bandit learning has been the subject of intense investigation over the past decade. The most natural categorization of these works is between parametric, realizability-based, and agnostic approaches. Parametric methods (cf. Abbasi-Yadkori et al. (2011); Chu et al. (2011)) typically assume a (generalized) linear relationship between the losses and the contexts/actions. Realizability-based methods generalize parametric methods by assuming the losses are predictable by some abstract regression class (Agarwal et al., 2012; Foster et al., 2018a). Agnostic approaches (cf. Auer et al. (2002b); Langford and Zhang (2008); Agarwal et al. (2014); Rakhlin and Sridharan (2016a); Syrgkanis et al. (2016)) avoid realizability assumptions and instead compete with VC-type policy classes for statistical tractability. Our work contributes to all of these directions, as our margin bounds apply to the agnostic adversarial setting and yield true regret bounds under realizability assumptions.

A special case of contextual bandits is *bandit multiclass prediction*, where the loss vector is zero for one action and one for all others (Kakade et al., 2008). A notable line of research derives surrogate regret bounds for this setting when the benchmark regressor class $\mathcal{F}$ consists of linear functions of the context (Kakade et al., 2008; Hazan and Kale, 2011; Beygelzimer et al., 2017; Foster et al., 2018b). Our work contributes to this line in two ways: we derive surrogate regret bounds and efficient algorithms beyond linear/parametric classes, and we consider the more general contextual bandit setting.

Our information-theoretic results on achievability are similar in spirit those of Daniely and Helbertal (2013), who derive tight generic bounds for bandit multiclass prediction in terms of the Littlestone dimension. This result is incomparable to our own: their bounds are on the $0/1$ loss regret directly rather than surrogate regret, but the Littlestone dimension is not a tight complexity measure for real-valued function classes in agnostic settings, which is the focusof the present work.

BIBLIOGRAPHY

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011. 296, 297, 351

Jacob Abernethy, Peter L. Bartlett, Alexaner Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 414–424. Omnipress, 2008. 38

Jacob D. Abernethy and Alexander Rakhlin. An efficient bandit algorithm for $O(\sqrt{T})$-regret in online multiclass prediction? In *Conference on Learning Theory*, 2009. 5, 16, 225, 227, 239, 240, 279, 281, 290, 296, 299

Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162 (3-4):531–586, 2015. 176

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, 2012. 296, 297, 351

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014. 36, 351

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Aleksandrs Slivkins. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016. 14, 279

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974. 7

Zeyuan Allen-Zhu and Yuanzhi Li. Follow the compressed leader: Faster algorithms for matrix multiplicative weight updates. *International Conference on Machine Learning*, 2017. 60

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009. 280

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012. 44, 59, 200

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017. 7, 14

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010. 7

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a. 7

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b. 41, 239, 283, 292, 351

Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 619–626. ACM, 2001. 43

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, June 2001. 75

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010. 229

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013. 229

Francis R Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014. 229

Raghu Raj Bahadur. Sufficiency and statistical decision functions. *Ann. Math. Statist*, 25(3): 423–462, 1954. 62

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 2

Keith Ball, Eric A Carlen, and Elliott H Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994. 199

Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *International Congress of Mathematicians*, 2014. 84

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003. ISSN 1532-4435. 8, 142, 190

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002. 109, 190

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006. 280, 282, 288

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017. 7

Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates. 2015. 7

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. 8

Shai Ben-David, David Pal, and Shai Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22th Annual Conference on Learning Theory*, 2009. 191

Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001. 43, 87

Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Comp. Linguistics*, 22(1), 1996. 225

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013. 10

Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357–365, 1944. 225

Quentin Berthet, Philippe Rigollet, et al. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013. 13

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011. 273

Alina Beygelzimer, Satyen Kale, and Haipeng Luo. Optimal and adaptive algorithms for online boosting. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2323–2331, 2015. 5, 16, 225, 227, 241, 245, 270

Alina Beygelzimer, Francesco Orabona, and Chicheng Zhang. Efficient Online Bandit Multiclass Learning with $\tilde{O}(\sqrt{T})$ Regret. In *ICML*, pages 488–497, 2017. 239, 290, 296, 351

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. 14

Daniel Billsus and Michael J Pazzani. Learning collaborative information filters. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 46–54. Morgan Kaufmann Publishers Inc., 1998. 42

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998. 109

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008. 3, 13

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005. 30, 31, 280, 283, 287

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. 47, 194, 207, 330, 345

George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953. 14

George EP Box and Norman R Draper. *Empirical model-building and response surfaces.* John Wiley & Sons, 1987. 14

Sebastien Bubeck, Nikhil Devanur, Zhiyi Huang, and Rad Niazadeh. Online auctions and multi-scale online learning. *The 18th ACM conference on Economics and Computation (EC 17)*, 2017. 222

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete and Computational Geometry*, 2018. 276, 294, 296, 331

Donald L. Burkholder. A geometrical characterization of banach spaces in which martingale difference sequences are unconditional. *The Annals of Probability*, pages 997–1011, 1981. 11, 48, 59, 67, 72

Donald L Burkholder. Boundary value problems and sharp inequalities for martingale transforms. *The Annals of Probability*, 12(3):647–702, 1984. 11, 48, 59, 141, 148, 150, 160, 165

Donald L Burkholder. Martingales and fourier analysis in banach spaces. In *Probability and analysis*, pages 61–108. Springer, 1986. 11, 48, 59, 150, 185

Donald L Burkholder. Explorations in martingale theory and its applications. In *École d'Été de Probabilités de Saint-Flour XIX?1989*, pages 1–66. Springer, 1991. 11, 48, 59

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007. 8

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. 43

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009. 3, 43

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006. 3, 8

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011. 8

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. 56, 62, 74, 100, 107, 112, 192, 230, 247, 258, 274, 335

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007. 112, 122

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, 2017. 279, 281, 289, 299, 349, 350

Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, page 201302293, 2013. 13

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6): 805–849, 2012. 8

Kamalika Chaudhuri, Yoav Freund, and Daniel J Hsu. A parameter-free hedging algorithm. In *Advances in neural information processing systems*, pages 297–305, 2009. 112, 121, 223

Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv:1805.01648*, 2018. 296

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, and Whitney K Newey. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016. 7, 14

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. 14

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, 2012. 112

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011. 288, 296, 297, 351

Thomas M. Cover. Behavior of sequential predictors of binary sequences. In *in Trans. 4th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, 1967. 19, 20

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958. 224

Sonja Cox and Mark Veraar. Some remarks on tangent martingale difference sequences in l1-spaces. *Electron. Comm. Probab*, 12(421-433):380, 2007. 160

Sonja Cox and Mark Veraar. Vector-valued decoupling and the burkholder–davis–gundy inequality. *Illinois Journal of Mathematics*, 55(1):343–375, 2011. 160

Ashok Cutkosky and Kwabena A Boahen. Online convex optimization with unconstrained domains and losses. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 748–756. 2016. 80, 196, 223

Ashok Cutkosky and Kwabena A. Boahen. Online learning without prior information. *The 30th Annual Conference on Learning Theory*, 2017. 80, 223

Ashok Cutkosky and Francesco Orabona. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. *Conference on Learning Theory*, 2018. 80, 222

Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *arXiv:1710.00095*, 2017. 296

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, 2013. 288, 351

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014. 193

Scott E Decatur, Oded Goldreich, and Dana Ron. Computational sample complexity. *SIAM Journal on Computing*, 29(3):854–879, 2000. 13

Luc Devroye, Lázló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. 189

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. 2

David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995. 3, 8

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 8, 73, 74, 112, 145, 157, 188

John C Duchi, Peter L. Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012. 294, 329

Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. 161

Eyal Even-Dar, Michael Kearns, Yishay Mansour, and Jennifer Wortman. Regret to the best vs. regret to the average. *Machine Learning*, 72(1-2):21–37, 2008. 121

R.A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368, 1922. 62

Dylan J. Foster and Akshay Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. *Advances in Neural Information Processing Systems*, 2018. 24

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems*, pages 3375–3383, 2015. 23, 107, 137, 151, 204, 223

Dylan J. Foster, Satyen Kale, Mehryar Mohri, and Karthik Sridharan. Parameter-free online learning via model selection. In *Advances in Neural Information Processing Systems 30*, pages 6020–6030, 2017a. 24, 80, 222

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Zigzag: A new approach to adaptive online learning. *30th Annual Conference on Learning Theory*, 2017b. 23, 24, 59, 60, 84, 187

Dylan J. Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. *International Conference on Machine Learning*, 2018a. 297, 351

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. *Conference on Learning Theory*, 2018b. 278, 290, 296, 350, 351

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Online learning: Sufficient statistics and the burkholder method. *Conference on Learning Theory*, 2018c. 23, 59, 107, 222

Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *24th Annual Conference on Learning Theory (COLT)*, 2011. 43, 57

Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156, 1996. 3

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 3, 244, 283

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001. 7

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013. 7

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, and Predag Klasnja. Action centered contextual bandits. In *Advances in neural information processing systems*, pages 5977–5985, 2017. 14, 279

Frank R Hampel, Elvezio Ronchetti, Peter J Rousseeuw, and Werner A Stahel. Robust statistics: the approach based on influence functions. 1986. 14

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014. 38

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. 14, 30

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. 43, 192, 214

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188, 2010. 112

Elad Hazan and Satyen Kale. Newtron: an efficient bandit algorithm for online multiclass prediction. In *Advances in Neural Information Processing Systems*, pages 891–899, 2011. 227, 239, 240, 296, 351

Elad Hazan and Nimrod Megiddo. Online learning with prior knowledge. In *Conference on Learning Theory*, 2007. 299

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007. 62, 74, 225, 232, 239, 256

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. In *Conference on Learning Theory*, pages 38–1, 2012. 43, 44, 59, 198, 220

Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory*, pages 197–209, 2014. 224, 225, 226, 229, 232, 233

David P. Helmbold and Manfred K. Warmuth. On weak learning. *J. Comput. Syst. Sci.*, 50 (3):551–573, 1995. 255

Paweł Hitczenko. Domination inequality for martingale transforms of a rademacher sequence. *Israel Journal of Mathematics*, 84(1-2):161–178, 1993. 160

Paweł Hitczenko. On a domination of sums of random variables by sums of conditionally independent ones. *The Annals of Probability*, pages 453–468, 1994. 160

Peter J Huber. Robust statistics. 1981. 14

Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*. Springer, 2016. 45, 73, 150, 152, 153, 159, 160, 166, 167, 178, 180, 181, 186

Young Hun Jung and Ambuj Tewari. Online boosting algorithms for multi-label ranking. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018. 245

Young Hun Jung, Jack Goetz, and Ambuj Tewari. Online multiclass boosting. In *Advances in Neural Information Processing Systems*, pages 920–929, 2017. 225, 227, 241, 242, 245, 270, 275

Sham M Kakade and Andrew Y Ng. Online bounds for bayesian algorithms. In *Advances in neural information processing systems*, pages 641–648, 2005. 233

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447. ACM, 2008. 227, 238, 279, 281, 290, 296, 351

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800. MIT Press, 2009a. 218

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009b. 239, 291

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(Jun):1865–1890, 2012. 199, 291

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994. 14, 30

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 8

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Information Theory*, 47(5):1902–1914, 2001. 190

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1155–1175, 2015. 112, 120, 121, 198, 223

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924, 2017. 38

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. 2017. 14

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012. 7

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008. 298, 351

Guillaume Lecué and Philippe Rigollet. Optimal learning with q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014. 31

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. 2

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006. 4, 28

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010. 14, 36, 279

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, pages 1260–1285, 2015. 110

Elliott H Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973. 55

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 2

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994. 62

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014. 108

László Lovász and Santosh Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 57–68. IEEE, 2006. 232, 276

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007. 232, 295, 296

Gábor Lugosi and Andrew B Nobel. Adaptive model selection using empirical complexities. *Annals of Statistics*, pages 1830–1864, 1999. 109

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pages 1286–1304, 2015. 112, 120, 121, 122, 223

Françoise Lust-Piquard and Gilles Pisier. Non commutative khintchine and paley inequalities. *Arkiv för matematik*, 29(1-2):241–260, 1991. 45, 53

Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, and Joel A Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014. 45, 53, 57

Colin L Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973. 7

Pascal Massart. *Concentration inequalities and model selection*, volume 10. Springer, 2007. 7, 110, 115, 190

David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. 10, 31, 109

H. Brendan McMahan and Jacob Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, pages 2724–2732, 2013. 8, 80, 188, 190, 196, 223

H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. *Proceedings of The 27th Conference on Learning Theory*, 2014. 77, 78, 80, 101, 112, 119, 188, 190, 196, 223

H. Brendan Mcmahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. In *Advances in neural information processing systems*, pages 2402–2410, 2012. 223

H. Brendan McMahan and Matthew Streeter. Open problem: Better bounds for online logistic regression. In *Conference on Learning Theory*, pages 44–1, 2012. 5, 16, 140, 224, 225, 226, 229, 231

Nishant A Mehta. Fast rates with high probability in exp-concave statistical learning. *International Conference on Artificial Intelligence and Statistics*, 2017. 226, 233, 256

Shahar Mendelson. Learning without Concentration. In *Conference on Learning Theory*, 2014. 110

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015. 2, 14

Hariharan Narayanan and Alexander Rakhlin. Efficient sampling from time-varying log-concave distributions. *Journal of Machine Learning Research*, 18:112:1–112:29, 2017. 276, 336

Fedor Nazarov and Sergei Treil. The hunt for a bellman function: applications to estimates for singular integral operators and to other classical problems of harmonic analysis. 1996. 59

Fedor Nazarov, Sergei Treil, and Alexander Volberg. Bellman function in stochastic control and harmonic analysis. In *Systems, approximation, singular integral operators, and related topics*, pages 393–423. Springer, 2001. 59

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012. 8

Arkadi Nemirovski. Topics in non-parametric statistics. *Ecole dEté de Probabilités de Saint-Flour*, 28:85, 2000. 31, 47

Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. 87, 208

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009. 7, 225

Arkadii Nemirovski, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983. 13, 43, 87, 188

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. 1998. 87

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *International Conference on Algorithmic Learning Theory*, 2013. 294, 339, 340

Jiazhong Nie, Wojciech Kotłowski, and Manfred K Warmuth. Online pca with optimal regrets. In *International Conference on Algorithmic Learning Theory*, pages 98–112. Springer, 2013. 199, 219, 220

Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014. 80, 188, 190, 196, 223

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 2016. 80, 188, 190, 196, 223

Adam Osekowski. Two inequalities for the first moments of a martingale, its square function and its maximal function. *Bulletin Polish Acad. Sci. Math.*, 53:441–449, 2005. 73, 74

Adam Osekowski. Sharp martingale and semimartingale inequalities. *Monografie Matematyczne*, 72, 2012. 59, 72, 73, 150, 152

Adam Osekowski. On the umd constant of the space ℓ_1^N. 2016. 157, 186

Adam Osekowski. Personal communication. 2017. 60

Dmitriy Panchenko. Some extensions of an inequality of vapnik and chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002. 7

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994. 26, 115

Bernardo Avila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, pages 1391–1399, 2013. 282

Gilles Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20:326–350, 1975. ISSN 0021-2172. 26, 50, 67, 68, 72, 291

Gilles Pisier. Martingales in banach spaces (in connection with type and cotype). course ihp, feb. 2–8, 2011. 2011. 163, 181, 196

David Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, CA, 1990. 17, 161

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, 2017. 296

Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction, 2012. Available at `http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf`. 122, 123, 178

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013. 112

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression. In *Conference on Learning Theory*, 2014. 110, 111, 112, 116, 187, 237, 238

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. *CoRR*, abs/1501.06598, 2015. URL `http://arxiv.org/abs/1501.06598`. 237, 238, 311

Alexander Rakhlin and Karthik Sridharan. BISTRO: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning*, 2016a. 290, 351

Alexander Rakhlin and Karthik Sridharan. A tutorial on online supervised learning with applications to node classification in social networks. *arXiv preprint arXiv:1608.09014*, 2016b. 23

Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. *Conference on Learning Theory*, 2017. 62, 291

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems 23*, pages 1984–1992, 2010. 22, 38, 108, 109, 111, 113, 161, 259, 262, 266, 267, 290, 304

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Beyond regret. *Journal of Machine Learning Research - Proceedings Track*, 19:559–594, 2011. 115

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. In *Advances in Neural Information Processing Systems 25*, pages 2150–2158, 2012. 107, 133, 194, 208

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 2014. 109, 191, 235, 236, 237, 238, 288, 289, 291, 306

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2): 111–153, 2015. 115, 116, 122, 126, 128, 269, 325

James Renegar. A polynomial-time algorithm, based on newton's method, for linear programming. *Mathematical Programming*, 40(1):59–93, 1988. 208

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005. 43

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 6

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000. 8

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012. 280

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997. 3, 7

Rocco A Servedio. Computational sample complexity and attribute-efficient learning. *Journal of Computer and System Sciences*, 60(1):161–178, 2000. 13

Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and fenchel duality. In *Advances in neural information processing systems*, pages 1265–1272, 2007. 225

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011. 3, 13

Shai Shalev-Shwartz, Ohad Shamir, and Eran Tromer. Using more data to speed-up training time. In *Artificial Intelligence and Statistics*, pages 1019–1027, 2012. 13

Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: learning, bounding, and transducing. *Journal of Machine Learning Research*, 15(1):3401–3423, 2014. 60

John Shawe-Taylor, Peter L. Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998. 10, 190

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016. 2, 14

Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958. 37

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Conference on Learning Theory*, 2011. 288

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012. 8

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005. 43

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010. 112

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in neural information processing systems*, pages 2645–2653, 2011. 69, 198, 290, 325

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015. 14

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E. Schapire. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, 2016. 351

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010. 38

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. 8

Ambuj Tewari and Susan A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, 2017. 14, 279

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. 58

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012. 45, 53, 54, 55

Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003. 31

Alexandre B Tsybakov. Introduction to nonparametric estimation. 2008. 29, 31

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975. 14

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 5, 11, 13, 29

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 4, 28

Tim Van Erven, Peter D Grünwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015. 229

Vladimir N. Vapnik. *Estimation of dependences based on empirical data*, volume 40. Springer-Verlag New York, 1982. 189

Vladimir N Vapnik. The nature of statistical learning theory. 1995. 14

Vladimir N. Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998. 7

Vladimir N. Vapnik and Alexey A. Chervonenkis. Algorithms with complete memory and recurrent algorithms in pattern recognition learning. *Automation and Remote Control*, (4): 606, 1968. 34, 41

Vladimir N. Vapnik and Alexey A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of Complexity*, 16(2):11, 1971. 5, 11, 17, 189

Vladimir Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990. 62

Vladimir Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60. ACM, 1995. 226, 229

Vladimir Vovk. Competitive on-line linear regression. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 364–370, Cambridge, MA, USA, 1998. MIT Press. 75

Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939. 5, 9, 29

Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear algebra and its applications*, 520:44–66, 2017. 176

David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996. 5

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017. 7

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004. 280

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014. 13

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003. 188, 225

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016. 8